编审者

主　编　明道绪　　（四川农业大学）
副主编　王钦德　　（山西农业大学）
　　耿社民　　（西北农林科技大学）
　　傅筑荫　　（贵州大学）

主　审　荣廷昭　　（四川农业大学）

参编者　崔　岘　　（甘肃农业大学）
　　陈文广　　（华南农业大学）
　　盛建华　　（河南农业大学）
　　郭春华　　（西南民族学院）
　　金　凤　　（内蒙古农业大学）
　　宋代军　　（西南农业大学）
　　刘学洪　　（云南农业大学）
　　张红平　　（四川农业大学）

第三版前言

《生物统计附试验设计》第三版是根据中国农业出版社“十五”高等农业院校教材出版规划组组织编写的，并纳入“面向21世纪课程教材”系列。第三版编委会由四川农业大学明道绪、山西农业大学王钦德、西北农林科技大学耿社民、贵州大学傅筑荫、甘肃农业大学崔　岘、华南农业大学陈文广、河南农业大学盛建华、西南民族学院郭春华、内蒙古农业大学金　凤、西南农业大学宋代军、云南农业大学刘学洪和四川农业大学张红平共12人组成，于2001年5月在四川农业大学召开了编写会议。在编写会议上全体编委认真讨论、审定了“编写大纲”，确定了章、节安排，内容取舍、深度、广度和详略；并进行了编写分工。初稿完成后，由主编明道绪教授负责统稿，进行了必要的修改与增删，并请主审——四川农业大学荣廷昭教授审阅。

本教材包括绪论（明道绪 编写），资料的整理（张红平 编写），平均数、标准差与变异系数（耿社民、金　凤编写），常用概率分布（金　凤、明道绪 编写），*t*检验（陈文广、郭春华编写），方差分析（王钦德、张红平 编写），次数资料分析—χ^2检验（耿社民、金凤 编写），直线回归与相关（郭春华 编写），多元线性回归与多项式回归（崔　岘 编写），协方差分析（宋代军 编写），非参数检验（刘学洪 编写），试验设计（傅筑荫、盛建华编写）共十二章（其中包含部分选用内容，用“*”注明），并附有常用生物统计方法的SAS程序（张红平 编写）以及常用统计数学用表。

在教材编写中力求做到科学性与实用性、先进性与针对性相统一；做到循序渐进、由浅入深、深入浅出、简明易懂；在正确阐述重要的统计学原理的同时，着重于基本概念、基本方法的介绍，特别注意学生动手能力的培养；每一种设计或分析方法都安排有步骤完整、过程详细的实例予以说明；各章都配备有习题（附简要答案）供读者练习。

全教材在保持本学科的系统性和科学性的前提下，注意引入本学科发展的新知识、新成果；注重拓宽学生的知识面和提高实践能力以及统计分析与计算机科学的结合；避免与交叉学科有关知识的重复，力求体现强基础、重应用和当前进行的素质教育和创新教育的教学目标。

本教材除可作为高等农业院校动物科学类专业教学用书外，也可作为水产养殖学、生物技术等专业开设《生物统计》课程的教学用书，对畜牧、水产、生物技术科技工作者亦有重要参考价值。

本教材在编写过程中参考了有关中外文献和专著，编者对这些文献和专著的作者、对热情指导、大力支持编写工作的中国农业出版社武旭峰同志以及为本教材统稿作了大量具体工作的张红平博士、承担绘图工作的邹祖银同志一并表示衷心的感谢！

限于编者水平，错误、缺点在所难免，敬请生物统计学专家和广大读者批评指正，以便再版时修改。

编　者

2001年12月18日

目　录

第一章　　绪　论

第一节　生物统计在畜禽、水产科学研究中的作用

为了推动畜牧业、水产业的发展，常常要进行科学研究。例如畜禽、水产品种资源研究，新品种的选育，新的饲养、管理技术研究等。这些研究都离不开调查或试验。进行调查或试验首先必须解决的问题是：如何合理地进行调查或试验设计。在实际研究工作中常常碰见这样的情况：由于调查或试验设计不合理，以至于无法从所获得的数据提取有用的信息，造成人力、物力和时间的浪费。若调查或试验设计方法好，用较少的人力、物力和时间即可收集到必要而有代表性的资料，从中获得可靠的结论，达到调查或试验的预期目的，收到事半功倍之效。

通过调查或试验能获得一定数量的数据。这些数据常常表现出程度不同的变异。例如测量100头猪的日增重所获得的100个数据，彼此不完全相同，表现出一定程度的变异；又如测量了200头黄牛的体高，所获得的200个数据，也表现出一定程度的变异。产生这种变异的原因，有的已被人们所了解。例如品种、性别、年龄、初始重、健康状况、饲养条件等不同，使得所测的猪的日增重、黄牛的体高表现出差异。另外还有许多内在和外在的因素还未被人们所认识。由于这些人们已了解的因素和人们尚未认识因而无法控制的因素的作用，使得通过调查或试验得来的数据普遍具有变异性。所以进行调查或试验还必须解决的第二个问题是：如何科学地整理、分析所收集得来的具有变异的资料，揭示出隐藏在其内部的规律性。合理地进行调查或试验设计、科学地整理、分析所收集得来的资料是生物统计（**Biometrics**）的根本任务。

生物统计是数理统计的原理和方法在生物科学研究中的应用，是一门应用数学。它在畜禽、水产科学研究中具有十分重要的作用。

一、提供试验或调查设计的方法

试验设计这一概念有广义与狭义之分，广义的试验设计是指试验研究课题设计，也就是指整个试验计划的拟定，包含课题名称、试验目的，研究依据、内容及预期达到的效果，试验方案，供试单位的选取、重复数的确定、试验单位的分组，试验的记录项目和要求，试验结果的分析方法，经济效益或社会效益的估计,已具备的条件,需要购置的仪器设备，参加研究人员的分工，试验时间、地点、进度安排和经费预算，成果鉴定，学术论文撰写等内容。狭义的试验设计主要是指试验单位(如动物试验的畜、禽)的选取、重复数目的确定及试验单位的分组。生物统计中的试验设计主要指狭义的试验设计。合理的试验设计能控制和降低试验误差，提高试验的精确性，为统计分析获得试验处理效应和试验误差的无偏估计提供必要的数据。

调查设计这一概念也有广义与狭义之分，广义的调查设计是指整个调查计划的制定，包括调查研究的目的、对象与范围，调查项目及调查表，抽样方法的选取，抽样单位、抽样数量的确定，数据处理方法，调查组织工作，调查报告撰写与要求，经费预算等内容。狭义的调查设计主要包含抽样方法的选取，抽样单位、抽样数目的确定等内容。生物统计中的调查

设计主要指狭义的调查设计。合理的调查设计能控制与降低抽样误差，提高调查的精确性，为获得总体参数的可靠估计提供必要的数据。

简而言之，试验或调查设计主要解决合理地收集必要而有代表性资料的问题。

二、提供整理、分析资料的方法

整理资料的基本方法是根据资料的特性将其整理成统计表、绘制成统计图。通过统计表、图可以大致看到所得资料集中、离散的情况。并利用所收集得来的数据计算出几个统计量，以表示该资料的数量特征、估计相应的总体参数。

统计分析最重要的内容是差异显著性检验。通过抽样调查或控制试验，获得的是具有变异的资料。产生变异的原因是什么？是由于进行比较的处理间，例如不同品种、不同饲料配方间有实质性的差异或是由于无法控制的偶然因素所引起？显著性检验的目的就在于承认并尽量排除这些无法控制的偶然因素的干扰，将处理间是否存在本质差异揭示出来。显著性检验的方法很多，常用的有t 检验——主要用于检验两个处理平均数差异是否显著；方差分析——主要用于检验多个处理平均数间差异是否显著；χ^2 检验——主要用于由质量性状得来的次数资料的显著性检验等。

统计分析的另一个重要内容是对试验指标或畜禽性状间的关系进行研究,或者研究它们之间的联系性质和程度，或者寻求它们之间的联系形式，即进行相关分析与回归分析。通过对资料进行相关、回归分析，可以揭示出试验指标或性状间的内在联系，为畜禽、水产新品种选育等提供强有力的依据。

还有一类统计分析方法不考虑资料的分布类型，也不事先对有关总体参数进行估算，这类统计分析方法叫非参数检验法。非参数检验法计算简便。当通常的检验方法对畜禽、水产科研中的某些资料无能为力时，非参数检验法则正好发挥作用。

以上我们对生物统计在畜禽、水产科学研究中的作用作了概略的介绍。从中不难看出，生物统计对于进行畜禽、水产科学研究是多么重要。它是每一个畜禽、水产科技工作者必须掌握的基本工具。可喜的是，随着生物统计方法的普及、计算工具的改进、统计计算程序的编制，已有越来越多的科技工作者掌握并在实际研究工作中应用了生物统计，取得了显著成效。

第二节　生物统计的常用术语

生物统计是一门应用数学，它涉及较多的数学概念、计算公式和数学用表；从判断方式上要求摆脱传统的确定性推断方式而接受建立在概率论基础上的统计推断方式,这对初学者来说有一定难度。为了便于初学者学习，在本教材中除了结合实例，从应用的角度来介绍生物统计的基本概念、基本原理、基本方法外，每章后还附有一定数量的习题供初学者练习。对于初学者来说，能正确理解生物统计的基本概念、了解基本原理、掌握并应用所介绍的基本的试验设计与结果分析方法解决畜牧、水产等科学研究中收集、整理、分析资料的问题，也就达到预期目的了。

在这一节里介绍生物统计中几个最常用的术语。

一、总体与样本

根据研究目的确定的研究对象的全体称为总体(**population**)，其中的一个研究单位称为个体(**individual**)；总体的一部分称为样本(**sample**)。例如研究中国黑白花乳牛头胎305天产乳量，所有中国黑白花乳牛头胎305天产乳量观测值的全体就构成中国黑白花乳牛头胎305天产乳量总体；而观测200头中国黑白花乳牛头胎305天产乳量所得的200个观测值则是中国黑白花乳牛头胎305天产乳量总体的一个样本，这个样本包含有200个个体。含有有限个个体的总体称为有限总体。例如上述中国黑白花乳牛头胎305天产乳量总体虽然包含的个体数目很多，但仍为有限总体。包含有无限多个个体的总体叫无限总体。例如在生物统计理论研究上的服从正态分布的总体、服从 t 分布的总体，包含一切实数，属于无限总体。在实际研究中还有一类假想总体。例如进行几种饲料的饲养试验，实际上并不存在用这几种饲料进行饲养的总体，只是假设有这样的总体存在，把所进行的试验看成是假想总体的一个样本。样本中所包含的个体数目叫样本容量或大小(**sample size**)。例如上述中国黑白花乳牛头胎305天产乳量样本容量为200。样本容量常记为n。通常把$n\leqslant30$的样本叫小样本，$n>30$的样本叫大样本。

生物统计一般是通过样本来了解总体。这是因为或者总体是无限的、假想的；即便是有限的但包含的个体数目相当多，要获得全部观测值须花费大量人力、物力和时间；或者观测值的获得带有破坏性，例如猪的瘦肉率测定，要求将猪屠宰后，把剥离板油和肾脏的胴体分割为瘦肉、脂肪、皮、骨四部分，再进行计算，不允许也没有必要对每一头猪一一屠宰测定。研究的目的是要了解总体，然而能观测到的却是样本，通过样本来推断总体是统计分析的基本特点。为了能可靠地从样本来推总体，要求样本具有一定的含量和代表性。只有从总体随机抽取的样本才具有代表性。所谓随机抽取**(random sampling)**是指总体中的每一个个体都有同等的机会被抽取组成样本。然而样本毕竟只是总体的一部分，尽管样本具有一定的含量也具有代表性，通过样本来推断总体也不可能是百分之百的正确。有很大的可靠性但有一定的错误率这是统计分析的又一特点。所以*Lienert*(1973)指出：作为科学方法论的现代统计学究竟能提供什么？它能回答在抽样调查中所发现的差异、联系和规律性以什么样的概率纯属偶然？对于总体来说这些发现作为一般规律的可靠程度有多大？

二、参数与统计量

为了表示总体和样本的数量特征，需要计算出几个特征数。由总体计算的特征数叫参数(**parameter**)；由样本计算的特征数叫统计量(**statistic**)。常用希腊字母表示参数，例如用μ表示总体平均数，用σ表示总体标准差；常用拉丁字母表示统计量，例如用 $\bar{x}$ 表示样本平均数，用S表示样本标准差。总体参数由相应的统计量来估计，例如用 $\bar{x}$ 估计μ，用S估计σ等。

三、准确性与精确性

准确性(**accuracy**)也叫准确度，指在调查或试验中某一试验指标或性状的观测值与其真值接近的程度。设某一试验指标或性状的真值为μ，观测值为 x，若 x与μ相差的绝对值$|x-\mu|$小，则观测值x的准确性高；反之则低。精确性(**precision**)也叫精确度，指调查或试验中同一试验指标或性状的重复观测值彼此接近的程度。若观测值彼此接近，即任意二个观测值x_i 、x_j 相差的绝对值$|x_i-x_j|$小，则观测值精确性高；反之则低。准确性、精确性的意

义图示如下：

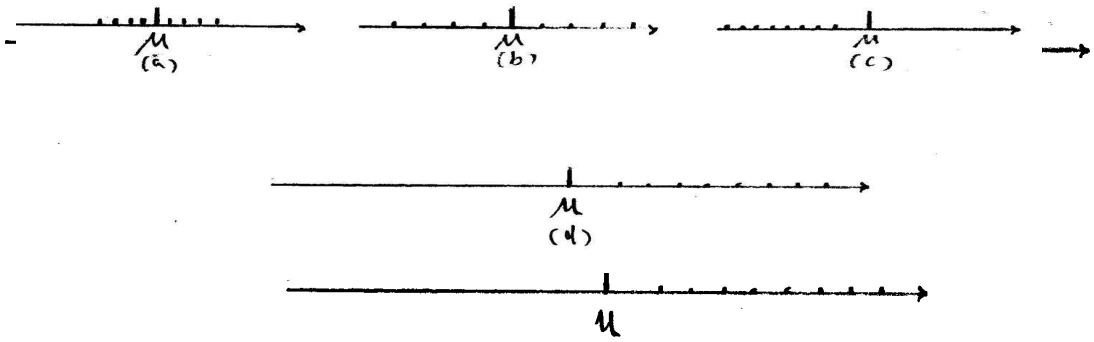


图1-1　准确性与精确性示

图1-1(*a*)观测值密集于真值μ两侧，其准确性高、精确性亦高；图1-1(*b*)观测值稀疏地分布于真值μ两侧，其准确性高，精确性却低；图1-1(*c*)观测值密集于远离真值μ的一侧，准确性低，精确性高；图1-1(*d*)观测值稀疏的分布于远离真值μ的一侧，其准确性、精确性都低。调查或试验的准确性、精确性合称为正确性。在调查或试验中应严格按照调查或试验计划进行，准确地进行观测记载，力求避免人为差错，特别要注意试验条件的一致性，即除所研究的各个处理外，供试畜禽的初始条件如品种、性别、年龄、健康状况、饲养条件、管理措施等应尽量控制一致，并通过合理的调查或试验设计努力提高试验的准确性和精确性。由于真值μ常常不知道，所以准确性不易度量，但利用统计方法可度量精确性。

四、随机误差与系统误差

在畜牧、水产科学试验中，试验指标除受试验因素影响外，还受到许多其它非试验因素的干扰，从而产生误差。试验中出现的误差分为两类：随机误差(**random error**)与系统误差(**systematic error**)。随机误差也叫抽样误差(**sampling error**)，这是由于许多无法控制的内在和外在的偶然因素如试验动物的初始条件、饲养条件、管理措施等尽管在试验中力求一致但不可能绝对一致所造成。随机误差带有偶然性质，在试验中，即使十分小心也难以消除。随机误差影响试验的精确性。统计上的试验误差指随机误差。这种误差愈小，试验的精确性愈高。系统误差也叫片面误差(**lopsided error**)，这是由于试验动物的初始条件如年龄、初始重、性别、健康状况等相差较大，饲料种类、品质、数量、饲养条件未控制相同，测量的仪器不准、标准试剂未经校正，以及观测、记载、抄录、计算中的错误所引起。系统误差影响试验的准确性。图1-1(*c*)、(*d*)所表示的情况，则是由于出现了系统误差的缘故。一般说来，只要试验工作做得精细，系统误差容易克服。图1-1(*a*)表示克服了系统误差的影响，且随机误差较小，因而准确性高，精确性也高。

第三节　　统计学发展概况

由于人类的统计实践是随着计数活动而产生的，因此，统计发展史可以追溯到远古的原始社会，也就是说距今足有五千多年的漫长岁月。但是，能使人类的统计实践上升到理论上予以概括总结的程度，即开始成为一门系统的学科统计学，却是近代的事情，距今只有三百余年的短暂历史。统计学发展的概貌，大致可划分为古典记录统计学、近代描述统计学和现代推断统计学三种形态。

一、古典记录统计学

古典记录统计学形成期间大致在十七世纪中叶至十九世纪中叶。统计学在这个兴起阶段，还是一门意义和范围不太明确的学问，在它用文字或数字如实记录与分析国家社会经济状况的过程中，初步建立了统计研究的方法和规则。到概率论被引进之后，才逐渐成为一项较成熟的方法。最初卓有成效地把古典概率论引进统计学的是法国天文学家、数学家、统计学家拉普拉斯(***P.S. Laplace***，1749～1827)。因此，后来比利时大统计学家凯特勒指出，统计学应从拉普拉斯开始。

（一）拉普拉斯的主要贡献

1、发展了概率论的研究　拉普拉斯第一种关于概率论的表述发表于1774年。从1812年起，先后出过四版《概率分析理论》，是他的代表作。书中，拉普拉斯最早系统地把数学分析方法运用到概率论研究中去，建立了严密的概率数学理论。该书不仅总结了他自己过去的研究，而且还总结了前一代学者研究概率论的成果，成为古典概率论的集大成者。

2、推广了概率论在统计中的应用　由于拉普拉斯是通过结合天文学、物理学的研究来从事概率研究的，所以，他能相当自觉、相当明确地指出：概率论能在广泛范围中应用，能解决一系列的实际问题。他在实际推广中的成绩是多方面的，主要表现在人口统计、观察误差理论和概率论对于天文问题的应用。1809～1812年，他结合概率分布模型和中心极限思想来研究最小二乘法，首次为统计学中这项后来最常用的手段奠定了理论基础。

3、明确了统计学的大数法则　拉普拉斯认为：“由于现象发生的原因，是为我们所不知或知道了也因为原因繁复而不能计算;发生原因又往往受偶然因素或无一定规律性因素所扰乱，以至事物发展发生的变化，只有进行长期大量观察，才能求得发展的真实规律。概率论则能研究此项发展改变原因所起作用的成份，并可指明成份多少。”这是他通过天文学上的研究后所得的体会。他发现在观察天体运动现象中，当次数足够多时，能使个体的特征趋于消失，而呈现出某种同一现象。他指出这其中一定存在着某些原因，而非出于偶然。

4、进行了大样本推断的尝试　在统计发展史上，人口的推算问题，多少年来成为统计学家耿耿于怀的难题。直到十九世纪初，拉普拉斯才用概率论的原理迈出了关键的一步。在理论上，1781年拉普拉斯在“论概率”一文中，建立了概率积分，为计算区间误差提供了有力手段。1781～1786年提出“拉普拉斯定理”(中心极限定理的一部分)，初步建立了大样本推断的理论基础。在实践上，拉普拉斯于1786年写了一篇关于巴黎人口的出生、婚姻、死亡的文章，文中提出根据法国特定地方的出生率来推算全国人口的问题。他抽选了30个市县，进行深入调查，推算出全国总人口数。尽管其方法和结果还相当粗糙，但在统计发展史上，他利用样本来推断总体的思想方法，为后人开创了一条抽样调查的新路子。

另一位对概率论与统计学的结合研究上作出贡献的是德国大数学家高斯(***C.F.Gauss***, 17

77～1855)。

（二）高斯的主要贡献

1、建立最小二乘法　在学生时代，高斯就开始了最小二乘法的研究。1794年，他读了数学家兰伯特(***J.H. Lambert***，1728～1777)的作品，讨论如何运用平均数法，从观察值(Y_i ，x_i)中确定线性关系Y＝α＋β*x*中的二个系数。1795年，设想了以残差平方和$\Sigma(Y_i - a - bx_i)^2$为最小的情况下，求得的*a*与*b*来估计α与β。1798年完成最小二乘法的整个思考结构，正式发表于1809年。

2、发现高斯分布　调查、观察或测量中的误差，不仅是不可避免的，而且一般是无法把握的。高斯以他丰富的天文观察和在1821～1825年间土地测量的经验，发现观察值***x***与真正值μ的误差变异，大量服从现代人们最熟悉的正态分布。他运用极大似然法及其他数学知识，推导出测量误差的概率分布公式。“误差分布曲线”这个术语就是高斯提出来的，后人为了纪念他，称这分布曲线为高斯分布曲线，也就是今天的正态分布曲线。高斯所发现的一般误差概率分布曲线以及据此来测定天文观察误差的方法，不仅在理论上，而且在应用上都有极重要的意义。

二、近代描述统计学

近代描述统计学形成期间大致在十九世纪中叶至二十世纪上半叶。由于这种“描述”特色由一批原是研究生物进化的学者们提炼而成，因此历史上称他们为生物统计学派。生物统计学派的创始人是英国的高尔登(***F. Galton***，1822～1911)，主将是高尔登的学生毕尔生(***K.Pearson***，1857～1936)。

（一） 高尔登的主要贡献

1、初创生物统计学　为了研究人类智能的遗传问题，高尔登仔细地阅读了三百多人的传记，以初步确定这些人中间多少人有亲属关系以及关系的大致密切程度。然后再从一组组知名人士中分别考察，以便从总体上来了解智力遗传的规律性。为了获得更多人的特性和能力的统计资料，高尔登自1882年起开设“人体测量实验室”。在连续六年中，共测量了9337人的“身高、体重、阔度、呼吸力、拉力和压力、手击的速率、听力、视力、色觉及个人的其它资料”，他深入钻研那些资料中隐藏着的内在联系，最终得出“祖先遗传法则”。他努力探索那些能把大量数据加以描述与比较的方法和途径，引入了中位数、百分位数、四分位数、四分位差以及分布、相关、回归等重要的统计学概念与方法。1901年，高尔登及其学生毕尔生在为《生物计量学》**(Biometrika)**杂志所写的创刊词中，首次为他们所运用的统计方法论明确提出了“生物统计”**(Biometry)**一词。高尔登解释道：“所谓生物统计学，是应用于生物学科中的现代统计方法”。从高尔登及后续者的研究实践来看，他们把生物统计学看作为一种应用统计学，其研究范围，既用统计方法来研究生物科学中的问题，更主要的是发展在生物科学应用中的统计方法本身。

2、对统计学的贡献

(1) 关于变异　变异是进化论中的重要概念，高尔登首次以统计方法加以处理，最终导致了英国生物统计学派的创立。1889年，高尔登把总体的定量测定法引入遗传研究中。高尔登通过总体测量发现，对动物或植物的每一个种别都可以决定一个平均类型。在一个种别中，

所有个体都围绕着这个平均类型，并把它当作轴心向多方面变异。这就是他在《遗传的天赋》一书中提出的“平均数离差法则”。

⑵关于“相关” 统计相关法是由高尔登创造的。关于相关研究的起因，最早是他因度量甜豌豆的大小，觉察到子代在遗传后有“返于中亲”的现象。1877年他搜集大量人体身长数据后，计算分析高个子父母、矮个子父母以及一高一矮父母的后代各有多少个高个子和矮个子子女，从而把父母高的后代高个子比较多、父母矮的其后代高个子比较少这一定性认识具体化为父母与子女之间在身长方面的定量关系。1888年，高尔登在“相关及其主要来自人体的度量”一文中，充分论述了“相关”的统计意义，并提出了高尔登相关函数(即现在常用的相关系数)的计算公式。

⑶ 关于“回归” 1870年，高尔登在研究人类身长的遗传时发现：高个子父母的子女，其身长有低于他们父母身长的趋势；相反，矮个子父母的子女，其身长却往往有高于他们父母身长的趋势，从人口全局来看，高个子的人“回归”于一般人身长的期望值，而矮个子的人则作相反的“回归”。这是统计学上“回归”的最初涵义。1886年，高尔登在论文“在遗传的身长中向中等身长的回归”中，正式提出了“回归”概念。

（二）毕尔生的主要贡献

对生物统计学倾注心血，并把它上升到通用方法论高度的是毕尔生。毕尔生的一生是统计研究的一生，他对统计学的主要贡献有：

1、变异数据的处理 生物统计中所取得的数据常常是零乱的，很难看出其所以然。为此，毕尔生首先探求处理数据的方法，他所首创的频数分布表与频数分布图如今已成为统计方法中最基本的手段之一。

2、分布曲线的选配 十九世纪以前，人们认为以频数分布描述变异值，最终都表现为正态分布曲线。但是，毕尔生从生物统计资料的经验分布中，注意到许多生物上的度量不具有正态分布，而常常呈偏态分布，甚至倾斜度很大；也不一定都是单峰，也有非单峰的。说明“唯正态”信念并不可靠。1894年，他在“关于不对称频率曲线的分解”一文中首先把非对称的观察曲线分解为几个正态曲线。他利用所谓“相对斜率”的方法得到12种分布函数型，其中包括正态分布、矩形分布、J 型分布、U型分布或铃型分布等。后来经R. 费雪的进一步研究，毕尔生分布曲线中第Ⅰ、Ⅱ、Ⅲ、Ⅳ及Ⅶ型出现在小样本理论内。尽管，毕尔生的曲线体系的推导方法是缺乏理论基础的，但也给人们不少启迪。

3、卡方检验的提出 1900年毕尔生独立地又重新发现了χ^2分布，并提出了有名的“卡方检验法”**(Test of χ^2)**。毕尔生获得了统计量：$\chi_q^2=\Sigma$(实际次数－理论次数)^2/理论次数，并证明了当观察次数充分大时，χ_q^2 总是近似地服从自由度为$(k-1)$的χ^2分布，其中k表示所划分的组数。在自然现象的范围内，χ^2检验法运用得很广泛。后经R. 费雪补充，成为了小样本推断统计的早期方法之一。

4、回归与相关的发展 回归与相关，经毕尔生进一步作了发展后，这两个出自于生物统计学领域的概念，便被推广为一般统计方法论的重要概念。 1896年，他在“进化论的数理研究： 回归、遗传和随机交配”一文中得出至今仍被广泛使用的线性相关计算公式：$\sum(x-\bar{x})(y-\bar{y})/\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}$ 。毕尔生还得出回归方程式：$\hat{y}=a+bx$ (其中a与b根据最小二乘法计算获得)，以及回归系数的计算公式：当y随x而变时，$\Sigma(x-\bar{x})(y-\bar{y})/$

$\Sigma(x-\bar{x})^2$；当x随y而变时，$\Sigma(x-\bar{x})(y-\bar{y})/\Sigma(y-\bar{y})^2$。此外，在1897～1905年，毕尔生还提出复相关、总相关、相关比等概念，不仅发展了高尔登的相关理论，还为之建立了数学基础。

三、现代推断统计学

现代推断统计学形成期间大致是二十世纪初叶至二十世纪中叶。人类历史进入二十世纪后，无论社会领域还是自然领域都向统计学提出更多的要求。各种事物与现象之间繁杂的数量关系以及一系列未知的数量变化，单靠记录或描述的统计方法已难以奏效。因此，相继产生“推断”的方法来掌握事物总体的真正联系以及预测未来的发展。从描述统计学到推断统计学，这是统计发展过程中的一个大飞跃。统计学发展中的这场深刻变革是在农业田间试验领域中完成的。因此，历史上称之为农业试验学派。对现代推断统计的建立贡献最大的是英国统计学家哥塞特(***W.S. Gosset***，1876～1937)和费雪(***R.A. Fisher***，1890～1962)。

（一） 哥塞特的*t*检验与小样本思想

1908年，哥塞特首次以“学生”(**Student**)为笔名，在《生物计量学》杂志上发表了“平均数的概率误差”。由于这篇文章提供了“学生t检验”的基础，为此，许多统计学家把1908年看作是统计推断理论发展史上的里程碑。后来，哥塞特又连续发表了“相关系数的概率误差”(1909)、“非随机抽样的样本平均数分布”(1909)、“从无限总体随机抽样平均数的概率估算表”(1917)，等等。他在这些论文中，第一，比较了平均误差与标准误差的两种计算方法；第二，研究了泊松分布应用中的样本误差问题；第三，建立了相关系数的抽样分布；第四，导入了“学生”分布，即*t*分布。这些论文的完成，为“小样本理论”奠定了基础；同时，也为以后的样本资料的统计分析与解释开创了一条崭新的路子。由于哥塞特开创的理论使统计学开始由大样本向小样本、由描述向推断发展，因此，有人把哥塞特推崇为推断统计学的先驱者。

（二） R. 费雪的统计理论与方法

R. 费雪一生先后共写作论文329篇。在世界各国流传最广泛的统计学著作是：1925年出版的《供研究人员用的统计方法》、1930年出版的《自然选择的遗传原理》、1935年出版的《试验设计》、1938年与耶特斯合著出版的《供生物学、农学与医学研究用的统计表》、1938年出版的《统计估计理论》、1950年出版的《对数理统计的贡献》、1950年出版的《统计方法和科学推断》等。当时，他在统计学方面居世界领先地位，他的贡献是多方面的。

1、“通用方法论” R. 费雪非常强调统计学是一门通用方法论，他认为无论对各种自然现象或社会生活现象的研究，统计方法及其计算公式“正如同其它数学科目一样，这里同一公式适用于一切问题的研究”。他指出“统计学是应用数学的最重要部分，并可以视为对观察得来的材料进行加工的数学”。

2、“假设无限总体” R. 费雪认为，在研究各种事物现象，包括社会经济现象时，必须把具体物质内容的信息舍弃掉，让统计处理的只是“统计总体”。比如说，“如果我们已有关于一万名新兵身长的资料，那么，统计研究的对象不是新兵的整体，而是各种身长尺寸的总体”。显然，R. 费雪只是对构成统计总体各因素的某些标志感兴趣而不是各因素的本身。其目的就是为了使问题简化，便于统计上的处理。他在1922年所写的“关于理论统计

学的数学基础”一文中，提出了一个重要的概念：“假设无限总体”。“所谓假设的无限总体，即现有的资料就是它的随机样本”。

3、抽样分布　R. 费雪跨进统计学界就是从研究概率分布开始的。1915年，他在《生物计量学》杂志上发表“无限总体样本相关系数值的频率分布”。由于这篇论文对相关系数的一般公式作了论证，对后来的整个推断统计的发展有一定贡献。因此，有人把这篇论文称为现代推断统计学的第一篇论文。1922年，R. 费雪导出相关系数***r***的***Z***分布，后来还编制了《Z曲线末端面积为0.05、0.01和0.001的Z数值分布表》。1924年，R. 费雪对t分布、χ^2分布和***Z***分布加以综合研究，使哥塞特的***t***检验也能适用于大样本，使毕尔生的χ^2检验也能适用于小样本。1938年，R. 费雪与耶特斯合编了《*F*分布显著性水平表》，为该分布的研究与应用，提供了方便。

4、方差分析　方差和方差分析两词，由R. 费雪于1918年在“孟德尔遗传试验设计间的相对关系”一文中所首创。方差分析也称变异数分析，其系统研究开始于1923年R. 费雪与麦凯基合写的“对收获量变化的研究”一文中。而于1925年，R. 费雪在《供研究人员用的统计方法》中对方差分析以及协方差分析进一步作了完整的叙述。“方差分析法是一种在若干能相互比较的资料组中，把产生变异的原因加以区分开来的方法与技术”。方差分析简单实用，大大提高了试验分析效率，对大样本、小样本都可使用。

5、试验设计　自1923年起，R. 费雪陆续发表了关于在农业试验中控制试验误差的论文。1925年他提出随机区组法和拉丁方法，到1926年，R. 费雪发表了试验设计方法的梗概；这些方法在1935年进一步得到完善，并首先在卢桑姆斯坦德农业试验站中得到检验与应用，后来又被他的学生推广到许多其它科学领域。

6、随机化原则　R. 费雪在创建试验设计理论的过程中，提出了十分重要的“随机化”原则。他认为这是保证取得无偏估计的有效措施，也是进行可靠的显著性检验的必要基础。所以，他把随机化原则放在极重要的地位，“要扫除可能扰乱资料的无数原因，除了随机化方法外，别无它法。”1938年，他和耶特斯合作编制了有名的*Fisher Yates*随机数字表。利用随机数字表保证总体中每一元素有同等被抽取的机会。这样，R. 费雪就把随机化原则以最明确、最具体化的形式引入统计工作与统计研究中。

R. 费雪在统计发展史上的地位是显赫的。这位多产作家的研究成果特别适用于农业与生物学领域，但它的影响已经渗透到一切应用统计学，由此所提炼出来的推断统计学已越来越被广大领域所接受。因此，美国统计学家约翰逊(*P.O. Johnson*)于1959年出版的《现代统计方法：描述和推断》一书中指出：“从1920年起一直到今天的这段时期，称之为统计学的费雪时代是恰当的”。

四、统计学在中国的传播

1913年，顾澄教授(1882～?)翻译了统计名著《统计学之理论》。这是英国统计学家尤尔在1911年新出版的关于描述统计学的著作，也就是英美数理统计学传入中国之始。之后有1922年翻译英国爱尔窦登的《统计学原理》、1929年翻译美国金氏的《统计方法》、1938年翻译鲍莱的《统计学原理》、1941年翻译密尔斯的《统计方法》。密尔斯的著作对中国统计学界影响较大，并被推崇为统计学范本。R. 费雪的理论和方法也很快传入中国，在20世

纪三十年代，“生物统计与田间试验”就作为农学系的必修课程，最早有1935年王绶编著出版的《实用生物统计法》，随后有范福仁著于1942年出版的《田间试验之设计与分析》。新中国成立后，中国科学院生物物理研究所的杨纪柯在介绍、推广数理统计学上作了大量工作。1963年他与汪安琦一起翻译出版了*G.W.* 斯奈迪格著《应用于农学和生物学试验的数理统计方法》，同年，他编写出版了《数理统计方法在医学科学中的应用》。接着，郭祖超的《医用数理统计方法》(1963)、范福仁的《田间试验技术》(1964)、《生物统计学》(1966)、赵仁熔的《大田作物田间试验统计方法》(1964)相继问世。到了七十年代，中国科学院数理研究所数理统计组先后出版了《常用数理统计方法》(1973)、《回归分析方法》(1974)、《方差分析》(1977)、《正交试验法》(1975)、《常用数理统计用表》(1974)。薛仲三的《医学统计方法和原理》(1978)、上海师范大学数学系概率统计研究组的《回归分析及其试验设计》(1978)等，这些都有力地推动了数理统计方法在中国的普及和应用。1978年12月，国家统计局在四川峨眉召开了统计教学、科研规划座谈会。会上明确提出“统计工作部门应该更好地运用数理统计方法”。这以后有关统计学的教材与论著如雨后春笋般涌现，如南京农业大学主编农业院校统编教材《田间试验和统计方法》(1979年第一版、1988年第二版)、贵州农学院主编农业院校统编教材《生物统计附试验设计》(1980年第一版、1989年第二版)，林德光编著的《生物统计的数学原理》(1982)、张尧庭、方开泰编著的《多元统计分析引论》(1982)、莫惠栋编著的《农业试验统计》(1988年第一版，1994年第二版)、明道绪主编的《兽医统计方法》(1991)、吴仲贤主编的《生物统计》(1994)、俞渭江、郭单元编著的《畜牧试验设计》(1995)等，译著有：杨纪珂、孙长鸣翻译出版的*R.G.D.* 斯蒂尔、*J.H.* 托里著的《数理统计的原理与方法　适用于生物科学》(1979)，关彦华、王平翻译[日]吉田实著《畜牧试验设计》(1984)等。随着计算机的迅速普及，统计电算程序SAS，SPSS等的引进，统计学在中国的应用与研究出现了崭新的局面。

习　　题

1、什么是生物统计？它在畜牧、水产科学研究中有何作用？

2、什么是总体、个体、样本、样本含量、随机样本？统计分析的两个特点是什么？

3、什么是参数、统计量？二者有何关系？

4、什么是试验的准确性与精确性？如何提高试验的准确性与精确性？

5、什么是随机误差与系统误差？如何控制、降低随机误差，避免系统误差？

6、统计学发展的概貌可分为哪几种形态？拉普拉斯、高斯、高尔敦、毕尔生、哥塞特、费雪各对统计学有何重要贡献？

第二章　资料的整理

由调查或试验收集来的原始资料，往往是零乱的，无规律性可循。只有通过统计整理，才能发现其内部的联系和规律性，从而揭示事物的本质。资料整理是进一步统计分析的基础，本章首先介绍资料的分类，然后介绍不同类型资料的整理方法。

第一节　资料的分类

正确地进行资料的分类是资料整理的前提。在调查或试验中，由观察、测量所得的数据按其性质的不同，一般可以分为数量性状资料、质量性状资料和半定量（等级）资料三大类。

一、数量性状资料

数量性状**(quantitative character)**是指能够以量测或计数的方式表示其特征的性状。观察测定数量性状而获得的数据就是数量性状资料**(data of quantitative characteristics)**。数量性状资料的记载有量测和计数两种方式，因而数量性状资料又分为计量资料和计数资料两种。

（一）计量资料　指用量测手段得到的数量性状资料，即用度、量、衡等计量工具直接测定的数量性状资料。其数据是用长度、容积、重量等来表示，如体高、产奶量、体重、绵羊剪毛量等。这种资料的各个观测值不一定是整数，两个相邻的整数间可以有带小数的任何数值出现，其小数位数的多少由度量工具的精度而定，它们之间的变异是连续性的。因此，计量资料也称为连续性变异资料。

（二）计数资料　指用计数方式得到的数量性状资料。在这类资料中，它的各个观察值只能以整数表示，在两个相邻整数间不得有任何带小数的数值出现。如猪的产仔数、鸡的产蛋数、鱼的尾数、母猪的乳头数等，这些观察值只能以整数来表示，各观察值是不连续的，因此该类资料也称为不连续性变异资料或间断性变异资料。

二、质量性状资料

质量性状**(qualitative character)**是指能观察到而不能直接测量的性状，如颜色、 性别、生死等。这类性状本身不能直接用数值表示，要获得这类性状的数据资料，须对其观察结果作数量化处理，其方法有以下两种：

（一）统计次数法　在一定的总体或样本中，根据某一质量性状的类别统计其次数，以次数作为质量性状的数据。例如，在研究猪的毛色遗传时，白猪与黑猪杂交，子二代中白猪、黑猪和花猪的头数分类统计如下表。

表2-1　白猪和黑猪子二代的毛色分离情况

毛色	次数（*f*）	频率（%）
白色	332	73.78
黑色	96	21.33
花色	22	4.89
合　计	450	100.00

这种由质量性状数量化得来的资料又叫次数资料。

（二）评分法 对某一质量性状，因其类别不同，分别给予评分。例如，在研究猪的肉色遗传时，常用的方法是将屠宰后2小时的猪眼肌横切面与标准图谱对比，由浅到深分别给予1—5分的评分，以便统计分析。

三、半定量（等级）资料

半定量或等级资料**(semi-quantitative or ranked data)**是指将观察单位按所考察的性状或指标的等级顺序分组，然后清点各组观察单位的次数而得的资料。这类资料既有次数资料的特点，又有程度或量的不同。如粪便潜血试验的阳性反应是在涂有粪便的棉签上加试剂后观察颜色出现的快慢及深浅程度分为六个等级；又如用某种药物治疗畜禽的某种疾病，疗效分为“无效”、“好转”、“显效”和“控制”四个级别；然后统计各级别的供试畜禽数。半定量资料在兽医研究中是常见的。

三种不同类型的资料相互间是有区别的，但有时可根据研究的目的和统计方法的要求将一种类型资料转化成另一种类型的资料。例如，兽医临床化验动物的白细胞总数得到的资料属于计数资料，根据化验的目的，可按白细胞总数正常或不正常分为两组，清点各组的次数，计数资料就转化为质量性状次数资料；如果按白细胞总数过高、正常、过低分为三组，清点各组次数，就转化成了半定量资料 。

第二节　　资料的整理

在对原始资料进行整理之前，首先要对全部资料进行检查与核对，然后再根据资料的类型及研究的目的对资料进行整理。

一、资料的检查与核对

检查和核对原始资料的目的在于确保原始资料的完整性和正确性。所谓完整性是指原始资料无遗缺或重复。所谓正确性是指原始资料的测量和记载无差错或未进行不合理的归并。检查中要特别注意特大、特小和异常数据（可结合专业知识作出判断）。对于有重复、异常或遗漏的资料，应予以删除或补齐；对有错误、相互矛盾的资料应进行更正，必要时进行复查或重新试验。资料的检查与核对工作虽然简单，但在统计处理工作中却是一项非常重要的步骤，因为只有完整、正确的资料，才能真实地反映出调查或试验的客观情况，才能经过统计分析得出正确的结论。

二、资料的整理方法

对原始资料进行检查核对后，根据资料中观测值的多少确定是否分组。当观测值不多($n \leqslant 30$)时，不必分组，直接进行统计分析。当观测值较多($n>30$)时，宜将观测值分成若干组，以便统计分析。将观测值分组后，制成次数分布表，即可看到资料的集中和变异情况。不同类型的资料，其整理的方法略有不同。

（一）计数资料的整理　现以50枚受精种蛋孵化出雏鸡的天数为例，说明计数料的整理。

表2-2　50枚受精种蛋孵化出雏鸡的天数

21	20	20	21	23	22	22	22	21	22	20	23	22	23	22	19	22	23
24	22	19	22	21	21	21	22	22	24	22	21	21	22	22	23	22	22
21	22	22	23	22	23	22	22	22	23	23	22	21	22				

小鸡出壳天数在19—24天范围内变动，有6个不同的观察值。用各个不同观察值进行分组，共分为6组，可得表2-3形式的次数分布表。

表2-3　50枚受精种蛋出雏天数的次数分布表

孵化天数	划线计数	次数（*f*）
19	𝍸	2
20	𝍸𝍸	3
21	𝍸𝍸	10
22	𝍸𝍸𝍸𝍸𝍸	24
23	𝍸𝍸	9
24	𝍸	2
合计		50

从表2-3可以看出：种蛋孵化出雏天数大多集中在21—23天，以22 天的最多，孵化天数较短（19—20天）和较长（24天）的都较少。

表2-4　100只蛋鸡每年产蛋数的次数分布表

产蛋数	划线计数	次数（*f*）
200—209	𝍸	2
210—219	𝍸𝍸	8
220—229	𝍸𝍸𝍸	15
230—239	𝍸𝍸𝍸𝍸	20
240—249	𝍸𝍸𝍸𝍸𝍸	23
250—259	𝍸𝍸𝍸𝍸	17
260—269	𝍸𝍸	8
270—279	𝍸𝍸	4
280—289	𝍸	2
290—299	𝍸	1
合 计		100

有些计数资料，观察值较多，变异范围较大，若以每一观察值为一组，则组数太多，而每组内包含的观察值太少，资料的规律性显示不出来。对于这样的资料，可扩大为以几个相邻观察值为一组，适当减少组数，这样资料的规律性就较明显，对资料进一步计算分析也

比较方便。例如观测某品种100只蛋鸡每年每只鸡产蛋数（原始资料略），其变异范围为200—299枚。这样的资料如以每个观察值为一组，则组数太多（该资料最多可分为100组），如间隔10枚为一组，则可使组数适当减少。经初步整理后分为10组，资料的规律性就比较明显，见表2-4。

从表2-4可以看到，大部分蛋鸡的年产蛋数在220—259枚，但也有少数蛋鸡每年产蛋数少到200—209枚，多到290—299枚。

(二)计量资料的整理 计量资料不能按计数资料的分组方法进行整理，在分组前需要确定全距、组数、组距、组中值及组限，然后将全部观测值划线计数归组。下面以126头基础母羊的体重资料为例， 说明其整理的方法及步骤。

【例2.1】 将126头基础母羊的体重资料(见表2-5)整理成次数分布表。

1、求全距 全距是资料中最大值与最小值之差，又称为极差(*range*)，用***R***表示，即

$$R=Max(x)-Min(x)$$

表2-5中，基础母羊的最大体重为65.0kg，最小体重为37.0kg，因此

$$R=65.0-37.0=28.0\text{kg}。$$

2、确定组数 组数的多少视样本含量及资料的变动范围大小而定，一般以达到既简化资料又不影响反映资料的规律性为原则。组数要适当，不宜过多，亦不宜过少。分组越多所求得的统计量越精确，但增大了运算量；若分组过少，资料的规律性就反映不出来，计算出的统计量的精确性也较差。一般组数的确定，可参考表2-6。

表2-5 126头基础母羊的体重资料

单位：*kg*

53.0	50.0	51.0	57.0	56.0	51.0	48.0	46.0	62.0	51.0	61.0	56.0	62.0	58.0	46.5
48.0	46.0	50.0	54.5	56.0	40.0	53.0	51.0	57.0	54.0	59.0	52.0	47.0	57.0	59.0
54.0	50.0	52.0	54.0	62.5	50.0	50.0	53.0	51.0	54.0	56.0	50.0	52.0	50.0	52.0
43.0	53.0	48.0	50.0	60.0	58.0	52.0	64.0	50.0	47.0	**37.0**	52.0	46.0	45.0	42.0
53.0	58.0	47.0	50.0	50.0	45.0	55.0	62.0	51.0	50.0	43.0	53.0	42.0	56.0	54.5
45.0	56.0	54.0	**65.0**	61.0	47.0	52.0	49.0	49.0	51.0	45.0	52.0	54.0	48.0	57.0
45.0	53.0	54.0	57.0	54.0	54.0	45.0	44.0	52.0	50.0	52.0	52.0	55.0	50.0	54.0
43.0	57.0	56.0	54.0	49.0	55.0	50.0	48.0	46.0	56.0	45.0	45.0	51.0	46.0	49.0
48.5	49.0	55.0	52.0	58.0	54.5									

表2-6 样本含量与组数

样本含量（*n*）	组 数
10—100	7—10
100—200	9—12
200—500	12—17
500以上	17—30

本例中，*n*＝126，根据表2-6，初步确定组数为10组。

3、确定组距 每组最大值与最小值之差称为组距，记为 *i*。分组时要求各组的组距

相等。组距的大小由全距与组数确定，计算公式为：

组距(i)＝全距／组数

本例i＝28.0／10≈3.0。

4、 确定组限及组中值　各组的最大值与最小值称为组限。最小值称为下限， 最大值称为上限。每一组的中点值称为组中值，它是该组的代表值。组中值与组限、组距的关系如下：

组中值＝(组下限＋组上限)/2＝组下限＋1/ 2组距＝组上限－1/2组距

由于相邻两组的组中值间的距离等于组距，所以当第一组的组中值确定以后，加上组距就是第二组的组中值，第二组的组中值加上组距就是第三组的组中值，其余类推。

组距确定后， 首先要选定第一组的组中值。在分组时为了避免第一组中观察值过多，一般第一组的组中值以接近于或等于资料中的最小值为好。第一组组中值确定后，该组组限即可确定，其余各组的组中值和组限也可相继确定。注意，最末一组的上限应大于资料中的最大值。

表2-5中， 最小值为37.0， 第一组的组中值取37.5，因组距已确定为3.0，所以

第一组的下限=37.5-(1/2)×3.0＝36.0；第一组的上限也就是第二组的下限为 36.0+3.0=39.0；第二组的上限也就是第三组的下限为39.0+3.0=42.0，……，以此类推，一直到某一组的上限大于资料中的最大值为止，于是可分组为：36.0—39.0，39.0—42.0，……。为了使恰好等于前一组上限和后一组下限的数据能确切归组，约定将其归入后一组。通常将上限略去不写。如第一组记为36.0 — ，第二组记为39.0 — ，……。

5、归组划线计数，作次数分布表　分组结束后，将资料中的每一观测值逐一归组，划线计数，然后制成次数分布表。如表2-5中，第一个观察值53.0，应归入表2-7中第六组，组限为51.0 — 54.0；第二个数50.0，应归入第五组，组限为48.0 — 51.0；依次将126个观察值都进行归组划线计数，制成次数分布表，见表2-7。

表2-7　　126头基础母羊的体重的次数分布表

组 别	组中值	划线计数	次数（f）
36.0—	37.5	正	1
39.0—	40.5	正	1
42.0—	43.5	正	6
45.0—	46.5	正正正	18
48.0—	49.5	正正正正	26
51.0—	52.5	正正正正	27
54.0—	55.5	正正正正	26
57.0—	58.5	正正	12
60.0—	61.5	正	7
63.0—	64.5	丅	2
合　计			126

次数分布表不仅便于观察资料的规律性，而且可根据它绘成次数分布图及计算平均数、标准差等统计量。从表2-7可以看出126头基础母羊体重资料分布的一般趋势：体重的变异范围在37.0—65.0kg，大部分母羊的体重在45.0—60.0kg之间。

在归组划线时应注意，不要重复或遗漏，归组划线后将各组的次数相加，结果应与样

本含量相等，如不等，证明归组划线有误，应予纠正。在分组后所得实际组数，有时和最初确定的组数不同，如第一组下限和资料中的最小值相差较大或实际组距比计算的组距为小，则实际分组的组数将比原定组数多；反之则少。

（三）质量性状资料、半定量（等级）资料的整理 对于质量性状资料、半定量（等级）资料，可按性状或等级进行分组，分别统计各组的次数，然后制成次数分布表。例如，研究山羊的角遗传时，用纯种的有角羊与无角羊交配，杂种一代全为无角羊，观察F_2代山羊共120只，有角无角的分离情况列于表2-8。

表2-8 F_2代山羊的有角无角分离情况

角	次 数（f）	频率（%）
无 角	87	72.50
有 角	33	27.50
合 计	120	100.00

又如，整理仔猪死亡情况资料可根据死亡原因将仔猪分组，并统计次数，计算出频率即构成比，见表2-9。

表2-9 仔猪死亡情况

死亡原因	死亡数	频率（%）
冻 死	15	19.23
发育不良	20	25.46
肺 炎	13	16.67
白 痢	10	12.82
寄生虫	20	25.64
合 计	78	100.00

第三节 常用统计表与统计图

统计表是用表格形式来表示数量关系；统计图是用几何图形来表示数量关系。用统计表与统计图，可以把研究对象的特征、内部构成、相互关系等简明、形象地表达出来，便于比较分析。

一、统计表

（一）统计表的结构和要求 统计表由标题、横标目、纵标目、线条、数字及合计构成，其基本格式如下表:

表号 标题

总横标目（或空白）	纵标目	合 计
横标目	数字资料	
合 计		

编制统计表的总原则：结构简单，层次分明，内容安排合理，重点突出，数据准确，便于理解和比较分析。具体要求如下：

1、标题　标题要简明扼要、准确地说明表的内容，有时须注明时间、地点。

2、标目　标目分横标目和纵标目两项。横标目列在表的左侧， 用以表示被说明事物的主要标志；纵标目列在表的上端，说明横标目各统计指标内容，并注明计算单位，如%、kg、cm等等。

3、数字　一律用阿拉伯数字，数字以小数点对齐，小数位数一致， 无数字的用“—”表示，数字是“0”的，则填写“0”。

4、线条　表的上下两条边线略粗，纵、横标目间及合计用细线分开， 表的左右边线可省去，表的左上角一般不用斜线。

（二）统计表的种类　统计表可根据纵、横标目是否有分组分为简单表和复合表两类。

1、简单表　由一组横标目和一组纵标目组成，纵横标目都未分组。 此类表适于简单资料的统计，如表2-10。

表2-10　某品种鸡杂种二代冠形分离情况

冠　形	次 数（*f*）	频率（%）
玫瑰冠	106	74.13
单　冠	37	25.87
合　计	143	100.00

2、复合表　由两组或两组以上的横标目与纵标目结合而成， 或由一组横标目与两组或两组以上的纵标目结合而成，或由两组或两组以上的横、纵标目结合而成。此类表适于复杂资料的统计，如表2-11。

表2-11　几种动物性食品的营养成分

品 别	百 分 比（%）					
	蛋白质	脂肪	糖类	无机盐	水分	其它
牛奶	3.3	4.0	5.0	0.7	87.0	——
牛肉	19.2	9.2	——	1.0	62.1	8.5
鸡蛋	11.9	9.3	1.2	0.9	65.5	11.2
咸带鱼	15.5	3.7	1.8	10.0	29.0	40.0

二、统计图

常用的统计图有长条图 **(bar chart)**、园图**(pie chart)**、线图**(linear chart)**、 直方图**(histogram)**和折线图**(broken-line chart)**等。图形的选择取决于资料的性质， 一般情况下，计量资料采用直方图和折线图，计数资料、质量性状资料、半定量（等级）资料常用长条图、线图或园图。

（一）统计图绘制的基本要求

1、标题简明扼要，列于图的下方。

2、纵、横两轴应有刻度，注明单位。

3、横轴由左至右、纵轴由下而上，数值由小到大； 图形长宽比例约5：4

或6：5。

4、图中需用不同颜色或线条代表不同事物时，应有图例说明。

（二）常用统计图及其绘制方法

1、长条图　它用等宽长条的长短或高低表示按某一研究指标划分属性种类或等级的次数或频率分布。如表示奶牛几种疾病的发病率；几种家畜对某一寄生虫感染的情况；不同公羊油汗色泽的次数分布情况等。如果只涉及一项指标，则采用单式长条图；如果涉及两个或两个以上的指标，则采用复式长条图。

在绘制长条图时，应注意以下几点：

（1）纵轴尺度从“0”开始，间隔相等，标明所表示指标的尺度及单位。

（2）横轴是长条图的共同基线，应标明各长条的内容。长条的宽度要相等，间隔相同。间隔的宽度可与长条宽度相同或者是其一半。

（3）在绘制复式长条图时，将同一属性种类、等级的两个或两个以上指标的长条绘制在一起，各长条所表示的指标用图例说明，同一属性种类、等级的各长条间不留间隔。

例如，根据表2-10绘制的长条图是单式的，见图2-1。根据表2-11绘制的长条图是复式的，见图2-2。

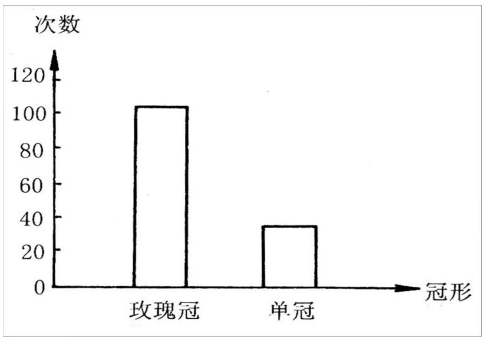


图 2-1 杂种二代鸡的冠形分离的次数分布图

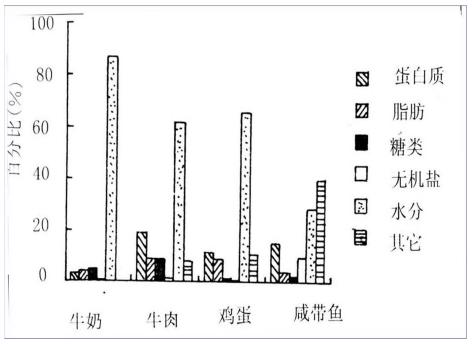


图 2-2　几种动物性食品的营养成分（条形

2、园图　用于表示计数资料、质量性状资料或半定量（等级）资料的构成比。 所谓构成比，就是各类别、等级的观测值个数(次数)与观测值总个数(样本含量)的百分比。把园图的全面积看成100%，按各类别、等级的构成比将园面积分成若干分， 以扇形面积的大小表分别表示各类别、等级的比例。

绘制园图时，应注意以下三点：

（1）圆图每3.6°园心角所对应的扇形面积为1%。

（2）圆图上各部分按资料顺序或大小顺序，以时钟9时或12时为起点，顺时针方向排列。

（3）圆图中各部分用线条分开，注明简要文字及百分比。

例如根据表2-11中的数据用圆图绘出四种动物性食品的营养成分，见图2-3。

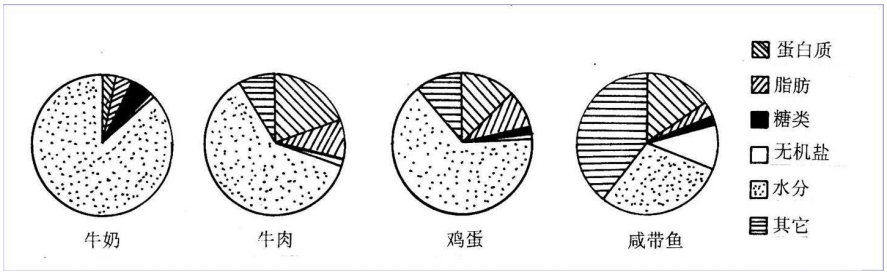


图 2-3 四种动物性食品的营养成分（圆图）

3、线图　用来表示事物或现象随时间而变化发展的情况。线图有单式和复式两种。

（1）单式线图　表示某一事物或现象的动态。

例如，某猪场长白猪从出生到6月龄出栏平均体重的变化如表2-12所示，根据该资料可以绘制成单式线图，以表示该猪场长白猪体重随月龄变化的情况，见图2-4。

表2-12　　长白猪体重的变化（出生——6月龄）

单位：*kg*

月龄	出生	1	2	3	4	5	6
体重	2.0	13.5	27.5	43.0	61.2	83.8	118.5

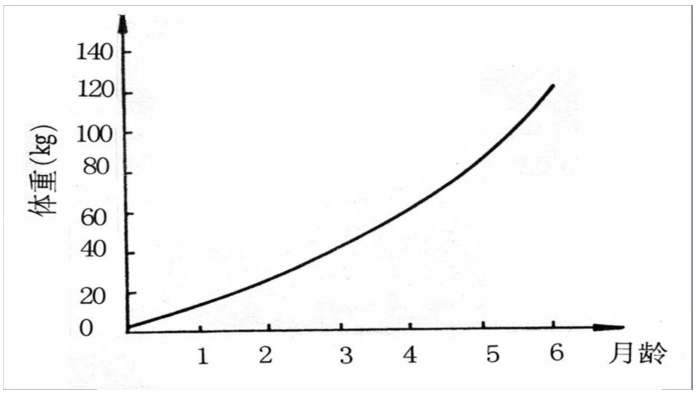


图2-4　长白猪体重的变化（0-6月龄）

（2）复式线图　在同一图上表示两种或两种以上事物或现象的动态。这时可用实线“——”，断线“------”，点线“• • • • •”，横点线“-• -• -• -”等来标志区别。

例如，长白猪、大约克、大白猪三个品种从出生到6月龄出栏平均体重的变化如表2-13所示，根据该资料绘制的复式线图，见图2-5。

表2-13　　三个品种猪体重的变化（出生——6月龄）

单位：*kg*

	出生	1	2	3	4	5	6
长白猪	2.0	13.5	27.5	43.0	61.2	83.8	118.5
大约克	1.8	12.0	24.5	38.0	53.6	72.3	104.5
大白猪	1.6	10.0	21.0	32.0	45.0	60.5	85.7

4、直方图(柱形图、矩形图) 对计量资料， 可根据次数分布表作出直方图以表示资料的分布情况。其作法是：在横轴上标记组限，纵轴标记次数（*f*），在各组上作出其高等于次数的矩形，即得次数分布直方图。

例如根据表2-7绘制的次数分布直方图，见图2-6。

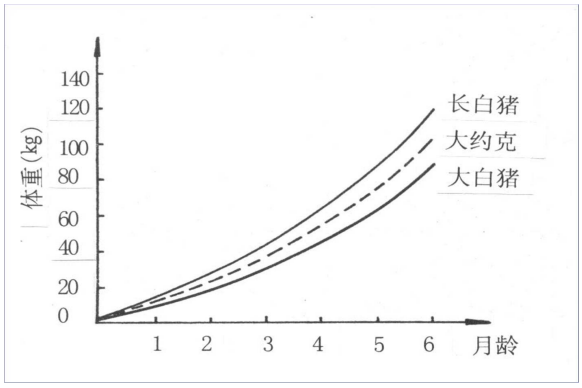


图 2-5 三个品种猪体重的变化（0-6 月龄）

5、折线图 对于计量资料，还可根据次数分布表作出次数分布折线图。 其作法是：在横轴上标记组中值，纵轴上标记次数，以各组组中值为横坐标，次数为纵坐标描点，用线段依次连接各点，即可得次数分布折线图。

例如根据表2-7绘制的次数分布折线图，见图2-7。

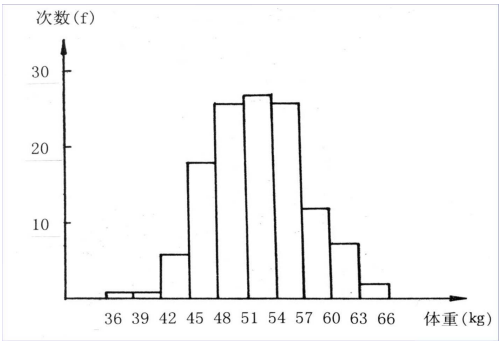


图 2-6 126 头基础母羊体重的次数分布直方

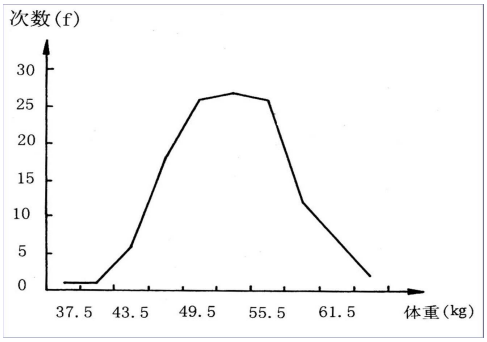


图 2-7 126 头基础母羊体重的次数分布折线图

习 题

1、 资料可以分为哪几类？它们有何区别与联系？

2、 为什么要对资料进行整理？对于计量资料，整理的基本步骤怎样？

3、 在对计量资料进行整理时，为什么第一组的组中值以接近或等于资料中的最小值为好？

4、 统计表与统计图有何用途？常用统计图有哪些？常用统计表有哪些？ 列统计表、绘统计图时，应注意什么？

5、 下表为100头某品种猪的血红蛋白含量（单位：g/100ml）资料，试将其整理成次数分布表，并绘制直方图和折线图。

13.4	13.8	14.4	14.7	14.8	14.4	13.9	13.0	13.0	12.8	12.5	12.3	12.1	11.8	11.0
10.1	11.1	10.1	11.6	12.0	12.0	12.7	12.6	13.4	13.5	13.5	14.0	15.0	15.1	14.1
13.5	13.5	13.2	12.7	12.8	16.3	12.1	11.7	11.2	10.5	10.5	11.3	11.8	12.2	12.4
12.8	12.8	13.3	13.6	14.1	14.5	15.2	15.3	14.6	14.2	13.7	13.4	12.9	12.9	12.4
12.3	11.9	11.1	10.7	10.8	11.4	11.5	12.2	12.1	12.8	9.5	12.3	12.5	12.7	13.0
13.1	13.9	14.2	14.9	12.4	13.1	12.5	12.7	12.0	12.4	11.6	11.5	10.9	11.1	11.6
12.6	13.2	13.8	14.1	14.7	15.6	15.7	14.7	14.0	13.9					

(提示：第一组下限取为9.1，组距i=0.7)

6、 测得某肉品的化学成分的百分比如下（单位：%），请绘制成圆图。

水 分	蛋白质	脂 肪	无机盐	其 它
62.0	15.3	17.2	1.8	3.7

7、 2001年调查四川省5个县奶牛的增长情况（与2000年相比）得如下资料（单位：%），请绘成长条图。

	双流县	名山县	宣汉县	青川县	泸定县
增长率（%）	22.6	13.8	18.2	31.3	9.5

8、 1-9周龄大型肉鸭杂交组合GW和GY的料肉比如下表所示，请绘制成线图。

周龄	1	2	3	4	5	6	7	8	9
GW	1.42	1.56	1.66	1.84	2.13	2.48	2.83	3.11	3.48
GY	1.47	1.71	1.80	1.97	2.31	2.91	3.02	3.29	3.57

第三章　平均数、标准差与变异系数

本章重点介绍平均数（**mean**）、标准差（**standard deviation**）与变异系数（**variation coefficient**）三个常用统计量，前者用于反映资料的集中性，即观测值以某一数值为中心而分布的性质；后两者用于反映资料的离散性，即观测值离中分散变异的性质。

第一节　平均数

平均数是统计学中最常用的统计量，用来表明资料中各观测值相对集中较多的中心位置。在畜牧业、水产业生产实践和科学研究中，平均数被广泛用来描述或比较各种技术措施的效果、畜禽某些数量性状的指标等等。平均数主要包括有算术平均数（**arithmetic mean**）、中位数（**median**）、众数（**mode**）、几何平均数（**geometric mean**）及调和平均数（**harmonic mean**），现分别介绍如下。

一、算术平均数

算术平均数是指资料中各观测值的总和除以观测值个数所得的商，简称平均数或均数，记为 $\bar{x}$。算术平均数可根据样本大小及分组情况而采用直接法或加权法计算。

（一）直接法　主要用于样本含量 $n \leqslant 30$ 以下、未经分组资料平均数的计算。

设某一资料包含 n 个观测值：x_1、x_2、…、x_n，则样本平均数 $\bar{x}$ 可通过下式计算：

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (3\text{-}1)$$

其中，Σ 为总和符号；$\sum_{i=1}^{n} x_i$ 表示从第一个观测值 x_1 累加到第 n 个观测值 x_n。当 $\sum_{i=1}^{n} x_i$ 在意义上已明确时，可简写为 Σx，（3-1）式即可改写为：

$$\bar{x} = \frac{\sum x}{n}$$

【例 3.1】 某种公牛站测得 10 头成年公牛的体重分别为 500、520、535、560、585、600、480、510、505、490（*kg*），求其平均体重。

由于 Σx=500+520+535+560+585+600+480+510+505+490=5285，n=10

代入（3—1）式得：

$$\bar{x} = \frac{\sum x}{n} = \frac{5285}{10} = 528.5(\text{kg})$$

即 10 头种公牛平均体重为 528.5 *kg*。

（二）加权法　对于样本含量 $n \geqslant 30$ 以上且已分组的资料，可以在次数分布表的基础上采用加权法计算平均数，计算公式为：

$$\bar{x}=\frac{f_1x_1+f_2x_2+\cdots+f_kx_k}{f_1+f_2+\cdots+f_k}=\frac{\sum_{i=1}^{k}f_ix_i}{\sum_{i=1}^{k}f_i}=\frac{\sum fx}{\sum f} \qquad (3\text{-}2)$$

式中：x_i—第 i 组的组中值；

f_i—第 i 组的次数；

k—分组数

第 i 组的次数 f_i 是权衡第 i 组组中值 x_i 在资料中所占比重大小的数量，因此 f_i 称为是 x_i 的“权”，加权法也由此而得名。

【例 3.2】 将 100 头长白母猪的仔猪一月窝重（单位：*kg*）资料整理成次数分布表如下，求其加权数平均数。

表 3—1　100 头长白母猪仔猪一月窝重次数分布表

组别	组中值（x）	次数（f）	fx
10—	15	3	45
20—	25	6	150
30—	35	26	910
40—	45	30	1350
50—	55	24	1320
60—	65	8	520
70—	75	3	225
合计		100	4520

利用（3—2）式得：

$$\bar{x}=\frac{\sum fx}{\sum f}=\frac{4520}{100}=45.2(kg)$$

即这 100 头长白母猪仔猪一月龄平均窝重为 45.2*kg*。

计算若干个来自同一总体的样本平均数的平均数时，如果样本含量不等，也应采用加权法计算。

【例 3.3】 某牛群有黑白花奶牛 1500 头，其平均体重为 750 *kg*，而另一牛群有黑白花奶牛 1200 头，平均体重为 725 *kg*，如果将这两个牛群混合在一起，其混合后平均体重为多少？

此例两个牛群所包含的牛的头数不等，要计算两个牛群混合后的平均体重，应以两个牛群牛的头数为权，求两个牛群平均体重的加权平均数，即

$$\bar{x}=\frac{\sum fx}{\sum f}=\frac{750\times1500+725\times1200}{2700}=738.89(kg)$$

即两个牛群混合后平均体重为 738.89 *kg*。

（三）平均数的基本性质

1、样本各观测值与平均数之差的和为零，即离均差之和等于零。

$$\sum_{i=1}^{n}(x_i-\bar{x})=0 \qquad \text{或简写成} \sum(x-\bar{x})=0$$

2、样本各观测值与平均数之差的平方和为最小，即离均差平方和为最小。

$$\sum_{i=1}^{n}(x_i-\bar{x})^2<\sum_{i=1}^{n}(x_i-a)^2 \qquad (\text{常数 } a\neq\bar{x})$$

或简写为：$\sum(x-\bar{x})^2<\sum(x-\alpha)^2$

以上两个性质可用代数方法予以证明，这里从略。

对于总体而言，通常用 μ 表示总体平均数，有限总体的平均数为：

$$\mu=\sum_{i=1}^{n}x_i \Big/ N \qquad (3\text{-}3)$$

式中，N 表示总体所包含的个体数。

当一个统计量的数学期望等于所估计的总体参数时，则称此统计量为该总体参数的无偏估计量。统计学中常用样本平均数（$\bar{x}$）作为总体平均数（μ）的估计量，并已证明样本平均数 $\bar{x}$ 是总体平均数 μ 的无偏估计量。

二、中位数

将资料内所有观测值从小到大依次排列，位于中间的那个观测值，称为中位数，记为 M_d。当观测值的个数是偶数时，则以中间两个观测值的平均数作为中位数。中位数简称中数。当所获得的数据资料呈偏态分布时，中位数的代表性优于算术平均数。中位数的计算方法因资料是否分组而有所不同。

（一）未分组资料中位数的计算方法 对于未分组资料，先将各观测值由小到大依次排列。

1、当观测值个数 n 为奇数时，$(n+1)/2$ 位置的观测值，即 $x_{(n+1)/2}$ 为中位数;

$$M_d=x_{(n+1)/2}$$

2、当观测值个数为偶数时，$n/2$ 和（$n/2+1$）位置的两个观测值之和的 1/2 为中位数，即:

$$M_d=\frac{x_{n/2}+x_{(n/2+1)}}{2} \qquad (3\text{-}4)$$

【例 3.4】 观察得 9 只西农莎能奶山羊的妊娠天数为 144、145、147、149、150、151、153、156、157，求其中位数。

此例 n=9，为奇数，则：

$$M_d=x_{(n+1)/2}=x_{(9+1)/2}=x_5=150\text{（天）}$$

即西农莎能奶山羊妊娠天数的中位数为 150 天。

【例 3.5】 某犬场发生犬瘟热，观察得 10 只仔犬发现症状到死亡分别为 7、8、8、9、11、12、12、13、14、14 天，求其中位数。

此例 n=10，为偶数，则：

$$M_d = \frac{x_{n/2} + x_{(n/2+1)}}{2} = \frac{x_5 + x_6}{2} = \frac{11+12}{2} = 11.5\text{（天）}$$

即 10 只仔犬从发现症状到死亡天数的中位数为 11.5 天。

（二）已分组资料中位数的计算方法 若资料已分组，编制成次数分布表，则可利用次数分布表来计算中位数，其计算公式为：

$$M_d = L + \frac{i}{f}(\frac{n}{2} - c) \qquad (3—5)$$

式中：*L*—中位数所在组的下限；

i—组距；

f—中位数所在组的次数；

n—总次数；

c—小于中数所在组的累加次数。

【例 3.6】 某奶牛场 68 头健康母牛从分娩到第一次发情间隔时间整理成次数分布表如表 3—2 所示，求中位数。

表 3—2 68 头母牛从分娩到第一次发情间隔时间次数分布表

间隔时间（*d*）	头数（*f*）	累加头数
12—26	1	1
27—41	2	3
42—56	13	16
57—71	20	36
72—86	16	52
87—101	12	64
102—116	2	66
≥117	2	68

由表 3—2 可见：*i*=15，*n*=68，因而中位数只能在累加头数为 36 所对应的“57—71”这一组，于是可确定 *L*=57，*f*=20，*C*=16，代入公式（3—5）得：

$$M_d = L + \frac{i}{f}(\frac{n}{2} - c) = 57 + \frac{15}{20}(\frac{68}{2} - 16) = 70.5\text{（天）}$$

即奶牛头胎分娩到第一次发情间隔时间的中位数为 70.5 天。

三、几何平均数

n 个观测值相乘之积开 *n* 次方所得的方根，称为几何平均数，记为 *G*。它主要应用于畜牧业、水产业的生产动态分析，畜禽疾病及药物效价的统计分析。如畜禽、水产养殖的增长率，抗体的滴度，药物的效价，畜禽疾病的潜伏期等，用几何平均数比用算术平均数更能代表其平均水平。其计算公式如下：

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n} = (x_1 \cdot x_2 \cdot x_3 \cdots x_n)^{\frac{1}{n}} \qquad (3—6)$$

为了计算方便，可将各观测值取对数后相加除以 *n*，得 *lgG*，再求 *lgG* 的反对数，即得

G 值，即

$$G = \lg^{-1}[\frac{1}{n}(\lg x_1 + \lg x_2 + \cdots + \lg x_n)] \qquad (3—7)$$

【例 3.7】 某波尔山羊群 1997—2000 年各年度的存栏数见表 3—3，试求其年平均增长率。

表 3—3　某波尔山羊群各年度存栏数与增长率

年度	存栏数（只）	增长率（x）	Lgx
1997	140	—	—
1998	200	0.429	-0.368
1999	280	0.400	-0.398
2000	350	0.250	-0.602
			Σlgx=-1.368

利用公式（3—7）求年平均增长率

$$G = \lg^{-1}[\frac{1}{n}(\lg x_1 + \lg x_2 + \cdots + \lg x_n)]$$

$$= lg^{-1}[\frac{1}{3}（\text{-0.368-0.398–0.602}）]$$

$$= lg^{-1}（\text{-0.456}）=0.3501$$

即年平均增长率为 0.3501 或 35.01%。

四、众　数

资料中出现次数最多的那个观测值或次数最多一组的组中值，称为众数，记为 M_0。如表 2-3 所列的 50 枚受精种蛋出雏天数次数分布中，以 22 出现的次数最多，则该资料的众数为 22 天。又如**【例 3.6】**所列出的次数分布表中，57—71 这一组次数最多，其组中值为 64 天，则该资料的众数为 64 天。

五、调和平均数

资料中各观测值倒数的算术平均数的倒数，称为调和平均数，记为 H，即

$$H = \frac{1}{\frac{1}{n}(\frac{1}{x_1} + \frac{1}{x_2} + \cdots \frac{1}{x_n})} = \frac{1}{\frac{1}{n}\sum\frac{1}{x}} \qquad (3—8)$$

调和平均数主要用于反映畜群不同阶段的平均增长率或畜群不同规模的平均规模。

【例 3.8】　某保种牛群不同世代牛群保种的规模分别为：0 世代 200 头，1 世代 220 头，2 世代 210 头；3 世代 190 头，4 世代 210 头，试求其平均规模。

利用公式（3—9）求平均规模：

$$H = \frac{1}{\frac{1}{5}(\frac{1}{200} + \frac{1}{220} + \frac{1}{210} + \frac{1}{190} + \frac{1}{210})} = \frac{1}{\frac{1}{5}(0.024)} = \frac{1}{0.0048} = 208.33（头）$$

即保种群平均规模为 208.33 头。

对于同一资料，算术平均数>几何平均数>调和平均数。

上述五种平均数，最常用的是算术平均数。

第二节　标准差

一、标准差的意义

用平均数作为样本的代表，其代表性的强弱受样本资料中各观测值变异程度的影响。如果各观测值变异小，则平均数对样本的代表性强；如果各观测值变异大，则平均数代表性弱。因而仅用平均数对一个资料的特征作统计描述是不全面的，还需引入一个表示资料中观测值变异程度大小的统计量。

全距（极差）是表示资料中各观测值变异程度大小最简便的统计量。全距大，则资料中各观测值变异程度大，全距小，则资料中各观测值变异程度小。但是全距只利用了资料中的最大值和最小值，并不能准确表达资料中各观测值的变异程度，比较粗略。当资料很多而又要迅速对资料的变异程度作出判断时，可以利用全距这个统计量。

为了准确地表示样本内各个观测值的变异程度，人们首先会考虑到以平均数为标准，求出各个观测值与平均数的离差，即（$x-\bar{x}$），称为离均差。虽然离均差能表达一个观测值偏离平均数的性质和程度，但因为离均差有正、有负，离均差之和为零，即 Σ（$x-\bar{x}$）=0，因而不能用离均差之和 Σ（$x-\bar{x}$）来表示资料中所有观测值的总偏离程度。为了解决离均差有正、有负，离均差之和为零的问题，可先求离均差的绝对值并将各离均差绝对值之和除以观测值 n 求得平均绝对离差，即 $\Sigma|x-\bar{x}|/n$。虽然平均绝对离差可以表示资料中各观测值的变异程度，但由于平均绝对离差包含绝对值符号，使用很不方便，在统计学中未被采用。我们还可以采用将离均差平方的办法来解决离均差有正、有负，离均差之和为零的问题。先将各个离均差平方，即 $(x-\bar{x})^2$，再求离均差平方和，即 $\Sigma(x-\bar{x})^2$，简称平方和，记为 *SS*；由于离差平方和常随样本大小而改变，为了消除样本大小的影响，用平方和除以样本大小，即 $\Sigma(x-\bar{x})^2/n$，求出离均差平方和的平均数；为了使所得的统计量是相应总体参数的无偏估计量，统计学证明，在求离均差平方和的平均数时，分母不用样本含量 n，而用自由度 n-1，于是，我们采用统计量 $\Sigma(x-\bar{x})^2/n-1$ 表示资料的变异程度。统计量 $\Sigma(x-\bar{x})^2/n-1$ 称为均方（**mean square** 缩写为 **MS**），又称样本方差，记为 S^2，即

$$S^2=\sum(x-\bar{x})^2/n-1 \qquad (3—9)$$

相应的总体参数叫总体方差，记为 σ^2。对于有限总体而言，σ^2 的计算公式为：

$$\sigma^2=\sum(x-\mu)^2/N \qquad (3—10)$$

由于样本方差带有原观测单位的平方单位，在仅表示一个资料中各观测值的变异程度而不作其它分析时，常需要与平均数配合使用，这时应将平方单位还原，即应求出样本方差的平方根。统计学上把样本方差 S^2 的平方根叫做样本标准差，记为 *S*，即：

$$S=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \tag{3-11}$$

由于 $\sum(x-\bar{x})^2=\sum(x^2-2x\bar{x}+\bar{x}^2)$

$$=\sum x^2-2\bar{x}\sum x+n\bar{x}^2$$

$$=\sum x^2-2\frac{(\sum x)^2}{n}+n(\frac{\sum x}{n})^2$$

$$=\sum x^2-\frac{(\sum x)^2}{n}$$

所以（3-11）式可改写为：

$$S=\sqrt{\frac{\sum x^2-\frac{(\sum x)^2}{n}}{n-1}} \tag{3-12}$$

相应的总体参数叫总体标准差，记为σ。对于有限总体而言，σ的计算公式为：

$$\sigma=\sqrt{\sum(x-\mu)^2/N} \tag{3-13}$$

在统计学中，常用样本标准差 S 估计总体标准差σ。

二、标准差的计算方法

（一）直接法 对于未分组或小样本资料，可直接利用（3—11）或（3-12）式来计算标准差。

【例 3.9】 计算 10 只辽宁绒山羊产绒量：450，450，500，500，500，550，550，550，600，600，650（g）的标准差。

此例 n=10，经计算得：Σx=5400，Σx^2=2955000，代入（3—12）式得：

$$S=\sqrt{\frac{\sum x^2-(\sum x)^2/n}{n-1}}=\sqrt{\frac{2955000-5400^2/10}{10-1}}=65.828\,(\text{g})$$

即 10 只辽宁绒山羊产绒量的标准差为 65.828g。

（二）加权法 对于已制成次数分布表的大样本资料，可利用次数分布表，采用加权法计算标准差。计算公式为：

$$S=\sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f-1}}=\sqrt{\frac{\sum fx^2-(\sum fx)^2/\sum f}{\sum f-1}} \tag{3—14}$$

式中，f 为各组次数；x 为各组的组中值；$\Sigma f=n$ 为总次数。

【例 3.10】 利用某纯系蛋鸡 200 枚蛋重资料的次数分布表（见表 3-4）计算标准差。

将表 3-4 中的 Σf、Σfx、Σfx^2 代入（3—14）式得：

$$S=\sqrt{\frac{\sum fx^2-(\sum fx)^2/\sum f}{\sum f-1}}=\sqrt{\frac{575507.11-10705.1^2/200}{200-1}}=3.5524\ （g）$$

即某纯系蛋鸡 200 枚蛋重的标准差为 3.5524*g*。

表 3—4　某纯系蛋鸡 200 枚蛋重资料次数分布及标准差计算表

组别	组中值（*x*）	次数（*f*）	*fx*	fx^2
44.15—	45.0	3	135.0	6075.0
45.85—	46.7	6	280.2	13085.34
47.55—	48.4	16	774.4	37480.96
49.25—	50.1	22	1102.2	55220.22
50.95—	51.8	30	1554.0	80497.20
52.65—	53.5	44	2354.0	125939.00
54.35—	55.2	28	1545.0	85317.12
56.05—	56.9	30	1707.0	97128.30
57.75—	58.6	12	703.2	41207.52
59.45—	60.3	5	301.5	18180.45
61.15—	62.0	4	248.0	15376.00
合计		Σ*f*=200	Σ*fx*=10705. 1	Σfx^2=575507. 11

三、标准差的特性

（一）标准差的大小，受资料中每个观测值的影响，如观测值间变异大，求得的标准差也大，反之则小。

（二）在计算标准差时，在各观测值加上或减去一个常数，其数值不变。

（三）当每个观测值乘以或除以一个常数 *a*，则所得的标准差是原来标准差的 *a* 倍或 1/*a* 倍。

（四）在资料服从正态分布的条件下，资料中约有 68.26%的观测值在平均数左右一倍标准差（$\bar{x}\pm S$）范围内；约有 95.43%的观测值在平均数左右两倍标准差（$\bar{x}\pm 2S$）范围内；约有 99.73%的观测值在平均数左右三倍标准差（$\bar{x}\pm 3S$）范围内。也就是说全距近似地等于 6 倍标准差，可用（全距/6）来粗略估计标准差。

第三节　变异系数

变异系数是衡量资料中各观测值变异程度的另一个统计量。当进行两个或多个资料变异程度的比较时，如果度量单位与平均数相同，可以直接利用标准差来比较。如果单位和（或）平均数不同时，比较其变异程度就不能采用标准差，而需采用标准差与平均数的比值（相对值）来比较。标准差与平均数的比值称为变异系数，记为 *C*·*V*。变异系数可以消除单位和

（或）平均数不同对两个或多个资料变异程度比较的影响。

变异系数的计算公式为：

$$C \cdot V = \frac{S}{\bar{x}} \times 100\% \qquad (3—15)$$

【例 3.11】 已知某良种猪场长白成年母猪平均体重为 190*kg*，标准差为 10.5*kg*，而大约克成年母猪平均体重为 196*kg*，标准差为 8.5*kg*，试问两个品种的成年母猪，那一个体重变异程度大。

此例观测值虽然都是体重，单位相同，但它们的平均数不相同，只能用变异系数来比较其变异程度的大小。

由于，长白成年母猪体重的变异系数：$C \cdot V = \frac{10.5}{190} \times 100\% = 5.53\%$

大约克成年母猪体重的变异系数：$C \cdot V = \frac{8.5}{196} \times 100\% = 4.34\%$

所以，长白成年母猪体重的变异程度大于大约克成年母猪。

注意，变异系数的大小，同时受平均数和标准差两个统计量的影响，因而在利用变异系数表示资料的变异程度时，最好将平均数和标准差也列出。

习　题

1、生物统计中常用的平均数有几种？各在什么情况下应用？

2、何谓算术平均数？算术平均数有哪些基本性质？

3、何谓标准差？标准差有哪些特性？

4、何谓变异系数？为什么变异系数要与平均数、标准差配合使用？

5、10 头母猪第一胎的产仔数分别为：9、8、7、10、12、10、11、14、8、9 头。试计算这 10 头母猪第一胎产仔数的平均数、标准差和变异系数。（$\bar{X}$=9.8 头，*S*=2.098 头，*C*・*V*=21.40%）。

6、随机测量了某品种 120 头 6 月龄母猪的体长，经整理得到如下次数分布表。试利用加权法计算其平均数、标准差与变异系数。

组别	组中值（*x*）	次数（*f*）
80—	84	2
88—	92	10
96—	100	29
104—	108	28
112—	116	20
120—	124	15
128—	132	13
136—	140	3

（$\bar{X}$=111.07cm，*S*=12.95cm, *C*・*V*=11.66%）。

7、某年某猪场发生猪瘟病，测得 10 头猪的潜伏期分别为 2、2、3、3、4、4、4、5、9、12(天)。试求潜伏期的中位数。（4 天）

8、某良种羊群 1995—2000 年六个年度分别为 240、320、360、400、420、450 只，试求该良种羊群的

年平均增长率。（G=0.1106 或 11.06%）。

9、某保种牛场，由于各方面原因使得保种牛群世代规模发生波动，连续 5 个世代的规模分别为：120、130、140、120、110 头。试计算平均世代规模。（H=123.17 头）

10、调查甲、乙两地某品种成年母水牛的体高（cm）如下表，试比较两地成年母水牛体高的变异程度。

甲地	137	133	130	128	127	119	136	132
乙地	128	130	129	130	131	132	129	130

（$S_{甲}$=5.75cm, $C.V_{甲}$=4.42%；$S_{乙}$=1.25cm，$C.V_{乙}$=0.96%）

第四章　常用概率分布

为了便于读者理解统计分析的基本原理，正确掌握和应用以后各章所介绍的统计分析方法， 本章在介绍概率论中最基本的两个概念——事件、概率的基础上，重点介绍生物科学研究中常用的几种随机变量的概率分布——正态分布、二项分布、波松分布以及样本平均数的抽样分布和 ***t*** 分布。

第一节　事件与概率

一、事　件

（一）必然现象与随机现象　在自然界与生产实践和科学试验中，人们会观察到各种各样的现象，把它们归纳起来，大体上分为两大类：一类是可预言其结果的，即在保持条件不变的情况下，重复进行试验，其结果总是确定的，必然发生（或必然不发生）。例如，在标准大气压下，水加热到 100℃必然沸腾；步行条件下必然不可能到达月球等。这类现象称为必然现象（**inevitable phenomena**）或确定性现象（**definite phenomena**）。另一类是事前不可预言其结果的，即在保持条件不变的情况下，重复进行试验，其结果未必相同。例如，掷一枚质地均匀对称的硬币，其结果可能是出现正面，也可能出现反面；孵化 6 枚种蛋，可能“孵化出 0 只雏”，也可能“孵化出 1 只雏”，…，也可能“孵化出 6 只雏”，事前不可能断言其孵化结果。这类在个别试验中其结果呈现偶然性、不确定性现象，称为随机现象（**random phenomena**）或不确定性现象（**indefinite phenomena**）。

人们通过长期的观察和实践并深入研究之后，发现随机现象或不确定性现象，有如下特点：在一定的条件实现时，有多种可能的结果发生，事前人们不能预言将出现哪种结果；对一次或少数几次观察或试验而言，其结果呈现偶然性、不确定性；但在相同条件下进行大量重复试验时，其试验结果却呈现出某种固有的特定的规律性——频率的稳定性，通常称之为随机现象的统计规律性。例如，对于一头临产的妊娠母牛产公犊还是产母犊是事前不能确定的，但随着妊娠母牛头数的增加，其产公犊、母犊的比例逐渐接近 1：1 的性别比例规律。概率论与数理统计就是研究和揭示随机现象统计规律的一门科学。

（二）随机试验与随机事件

1、随机试验　通常我们把根据某一研究目的，在一定条件下对自然现象所进行的观察或试验统称为试验（**trial**）。而一个试验如果满足下述三个特性，则称其为一个随机试验（**random trial**），简称试验：

（1）试验可以在相同条件下多次重复进行；

（2）每次试验的可能结果不止一个，并且事先知道会有哪些可能的结果；

（3）每次试验总是恰好出现这些可能结果中的一个，但在一次试验之前却不能肯定这次试验会出现哪一个结果。

如在一定孵化条件下，孵化 6 枚种蛋，观察其出雏情况；又如观察两头临产妊娠母牛所

产犊牛的性别情况，它们都具有随机试验的三个特征，因此都是随机试验。

2、随机事件　随机试验的每一种可能结果，在一定条件下可能发生，也可能不发生，称为随机事件（**random event**），简称事件（**event**），通常用 *A*、*B*、*C* 等来表示。

（1）基本事件　我们把不能再分的事件称为基本事件（**elementary event**），也称为样本点（**sample point**）。例如，在编号为 1、2、3、…、10 的十头猪中随机抽取 1 头，有 10 种不同的可能结果：“取得一个编号是 1”、“取得一个编号是 2”、…、“取得一个编号是 10”，这 10 个事件都是不可能再分的事件，它们都是基本事件。由若干个基本事件组合而成的事件称为复合事件（compound event）。如“取得一个编号是 2 的倍数”是一个复合事件，它由“取得一个编号是 2”、“是 4”、“是 6、“是 8”、“是 10”5 个基本事件组合而成。

（2）必然事件　我们把在一定条件下必然会发生的事件称为必然事件（**certain event**），用Ω表示。例如，在严格按妊娠期母猪饲养管理的要求饲养的条件下，妊娠正常的母猪经 114 天左右产仔，就是一个必然事件。

（3）不可能事件　我们把在一定条件下不可能发生的事件称为不可能事件（**impossible event**），用Φ表示。例如，在满足一定孵化条件下，从石头孵化出雏鸡，就是一个不可能事件。

必然事件与不可能事件实际上是确定性现象，即它们不是随机事件，但是为了方便起见，我们把它们看作为两个特殊的随机事件。

二 、 概　率

（一）概率的统计定义　研究随机试验，仅知道可能发生哪些随机事件是不够的，还需了解各种随机事件发生的可能性大小，以揭示这些事件的内在的统计规律性，从而指导实践。这就要求有一个能够刻划事件发生可能性大小的数量指标，这指标应该是事件本身所固有的，且不随人的主观意志而改变，人们称之为概率（**probability**）。事件 ***A*** 的概率记为 *P*（*A*）。下面我们先介绍概率的统计定义。

在相同条件下进行 *n* 次重复试验，如果随机事件 ***A*** 发生的次数为 *m*，那么 *m/n* 称为随机事件 ***A*** 的频率（**frequency**）；当试验重复数 *n* 逐渐增大时，随机事件 ***A*** 的频率越来越稳定地接近某一数值 *p*，那么就把 *p* 称为随机事件 ***A*** 的概率。这样定义的概率称为统计概率（**statistics probability**），或者称后验概率（**posterior probability**）。

例如为了确定抛掷一枚硬币发生正面朝上这个事件的概率，历史上有人作过成千上万次抛掷硬币的试验。在表 4—1 中列出了他们的试验记录。

表 4—1 抛掷一枚硬币发生正面朝上的试验记录

实验者	投掷次数	发生正面朝上的次数	频率（*m/n*）
蒲　丰	4040	2048	0.5069
k．皮尔逊	12000	6019	0.5016
k．皮尔逊	24000	12012	0.5005

从表 4-1 可看出，随着实验次数的增多，正面朝上这个事件发生的频率越来越稳定地接近 0.5，我们就把 0.5 作为这个事件的概率。

在一般情况下，随机事件的概率 *p* 是不可能准确得到的。通常以试验次数 *n* 充分大时随机事件 ***A*** 的频率作为该随机事件概率的近似值。

即　　$P(A)=p\approx m/n$　（n 充分大）　　（4-1）

（二）概率的古典定义　上面介绍了概率的统计定义。但对于某些随机事件，用不着进行多次重复试验来确定其概率，而是根据随机事件本身的特性直接计算其概率。

有很多随机试验具有以下特征：

1、试验的所有可能结果只有有限个，即样本空间中的基本事件只有有限个；

2、各个试验的可能结果出现的可能性相等，即所有基本事件的发生是等可能的；

3、试验的所有可能结果两两互不相容。

具有上述特征的随机试验，称为古典概型（**classical model**）。对于古典概型，概率的定义如下：

设样本空间由 n 个等可能的基本事件所构成，其中事件 A 包含有 m 个基本事件，则事件 A 的概率为 m/n，即

$$P(A)=m/n \qquad (4\text{-}2)$$

这样定义的概率称为古典概率(**classical probability**)或先验概率(**prior probability**)。

【例 4.1】在编号为 1、2、3、…、10 的十头猪中随机抽取 1 头，求下列随机事件的概率。

（1）A=“抽得一个编号≤4”；

（2）B=“抽得一个编号是 2 的倍数”。

因为该试验样本空间由 10 个等可能的基本事件构成，即 n=10,而事件 A 所包含的基本事件有 4 个，既抽得编号为 1，2，3，4 中的任何一个，事件 A 便发生，即 m_A=4，所以

$$P(A)=m_A/n=4/10=0.4$$

同理，事件 B 所包含的基本事件数 m_B=5，即抽得编号为 2，4，6，8，10 中的任何一个，事件 B 便发生，故 P(B)=m_B/n=5/10=0.5。

【例 4.2】　在 N 头奶牛中，有 M 头曾有流产史，从这群奶牛中任意抽出 n 头奶牛，试求：

(1)其中恰有 m 头有流产史奶牛的概率是多少？

(2)若 N=30，M =8，n =10，m =2，其概率是多少？

我们把从有 M 头奶牛曾有流产史的 N 头奶牛中任意抽出 n 头奶牛，其中恰有 m 头有流产史这一事件记为 A，因为从 N 头奶牛中任意抽出 n 头奶牛的基本事件总数为 C_N^n，事件 A 所包含的基本事件数为 $C_M^m \cdot C_{N-M}^{n-m}$ ，因此所求事件 A 的概率为

$$P(A)=\frac{C_M^m \cdot C_{N-M}^{n-m}}{C_N^n}$$

将 N=30，M =8，n =10，m =2 代入上式，得

$$P(A)=\frac{C_8^2 \cdot C_{30-8}^{10-2}}{C_{30}^{10}}=0.0695$$

即在 30 头奶牛中有 8 头曾有流产史，从这群奶牛随机抽出 10 头奶牛其中有 2 头曾有流产史的概率为 6.95%。

（三）概率的性质　根据概率的定义，概率有如下基本性质：

1、对于任何事件 A，有 $0\leqslant P(A)\leqslant 1$；

2、必然事件的概率为 1，即 $P(\Omega)=1$；

3、不可能事件的概率为 0，即 P（ ϕ ）=0。

三、小概率事件实际不可能性原理

随机事件的概率表示了随机事件在一次试验中出现的可能性大小。若随机事件的概率很小，例如小于 0.05、0.01、0.001，称之为小概率事件。小概率事件虽然不是不可能事件，但在一次试验中出现的可能性很小，不出现的可能性很大，以至于实际上可以看成是不可能发生的。在统计学上，把小概率事件在一次试验中看成是实际不可能发生的事件称为小概率事件实际不可能性原理，亦称为小概率原理。小概率事件实际不可能性原理是统计学上进行假设检验（显著性检验）的基本依据。在下一章介绍显著性检验的基本原理时，将详细叙述小概率事件实际不可能性原理的具体应用。

第二节　概率分布

事件的概率表示了一次试验某一个结果发生的可能性大小。若要全面了解试验，则必须知道试验的全部可能结果及各种可能结果发生的概率，即必须知道随机试验的概率分布**(probability distribution)**。为了深入研究随机试验，我们先引入随机变量**(random variable)**的概念。

一、随机变量

作一次试验，其结果有多种可能。每一种可能结果都可用一个数来表示，把这些数作为变量x的取值范围，则试验结果可用变量x来表示。

【例4.3】　对100头病畜用某种药物进行治疗，其可能结果是“0头治愈”、 “1头治愈”、“2头治愈”、“…”、“100头治愈”。若用x表示治愈头数，则x的取值为0、1、2、…、100。

【例4.4】　孵化一枚种蛋可能结果只有两种，即“孵出小鸡”与“未孵出小鸡”。 若用变量x表示试验的两种结果，则可令x=0表示“未孵出小鸡”，x=1表示“孵出小鸡”。

【例4.5】　测定某品种猪初生重，表示测定结果的变量x所取的值为一个特定范围(a,b)，如0.5－1.5kg，x值可以是这个范围内的任何实数。

如果表示试验结果的变量x，其可能取值至多为可列个，且以各种确定的概率取这些不同的值，则称x为离散型随机变量**(discrete random variable)**；如果表示试验结果的变量x，其可能取值为某范围内的任何数值，且x在其取值范围内的任一区间中取值时，其概率是确定的，则称x为连续型随机变量**(continuous random variable)**。

引入随机变量的概念后，对随机试验的概率分布的研究就转为对随机变量概率分布的研究了。

二、离散型随机变量的概率分布

要了解离散型随机变量x的统计规律，就必须知道它的一切可能值x_i及取每种可能值的概率p_i。

如果我们将离散型随机变量x的一切可能取值x_i (i=1, 2, …)，及其对应的概率p_i，记作

$$P(x=x_i)=p_i \qquad i=1, 2, \cdots \qquad (4—3)$$

则称(4—3)式为离散型随机变量x的概率分布或分布。常用分布列(**distribution series**)来表示离散型随机变量：

$$\begin{bmatrix} x_1 & x_2 & \dots & x_n & \dots \\ p_1 & p_2 & \dots & p_n & \dots \end{bmatrix}$$

显然离散型随机变量的概率分布具有$p_i \geqslant 0$和$\Sigma p_i=1$这两个基本性质。

三、连续型随机变量的概率分布

连续型随机变量(如体长、体重、蛋重)的概率分布不能用分布列来表示，因为其可能取的值是不可数的。我们改用随机变量x在某个区间内取值的概率P($a \leqslant x<b$)来表示。 下面通过频率分布密度曲线予以说明。

由表2—7作126头基础母羊体重资料的频率分布直方图，见图4—1，图中纵座标取频率与组距的比值。可以设想，如果样本取得越来越大($n \to +\infty$)，组分得越来越细($i \to 0$)，某一范围内的频率将趋近于一个稳定值——概率。这时，频率分布直方图各个直方上端中点的联线——频率分布折线将逐渐趋向于一条曲线，换句话说，当$n \to +\infty$、$i \to 0$时，频率分布折线的极限是一条稳定的函数曲线。 对于样本是取自连续型随机变量的情况，这条函数曲线将是光滑的。 这条曲线排除了抽样和测量的误差，完全反映了基础母羊体重的变动规律。 这条曲线叫概率分布密度曲线，相应的函数叫概率分布密度函数。若记体重概率分布密度函数为$f(x)$，则x取值于区间[a,b）的概率为图中阴影部分的面积，即

$$P(a \leqslant x<b)=\int_a^b f(x)dx \qquad (4\text{-}4)$$

(4—4)式为连续型随机变量x在区间[a,b）上取值概率的表达式。可见，连续型随机变量的概率由概率分布密度函数确定。

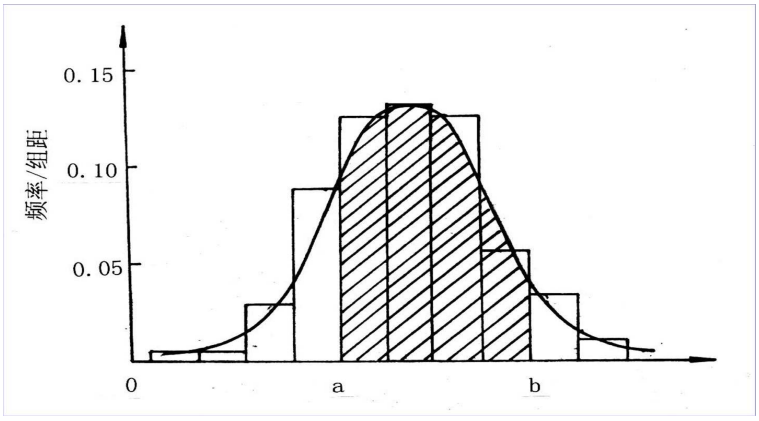


图4-1　表2-7资料的分布曲线

此外，连续型随机变量概率分布还具有以下性质：

1、分布密度函数总是大于或等于0，即$f(x) \geqslant 0$；

2、当随机变量x取某一特定值时，其概率等于0；即

$$P(x = c) = \int_c^c f(x)dx = 0 \qquad (c\text{为任意实数})$$

因而，对于连续型随机变量，仅研究其在某一个区间内取值的概率，而不去讨论取某一个值的概率。

3、在一次试验中随机变量x之取值必在$-\infty < x < +\infty$范围内，为一必然事件。所以

$$P(-\infty < x < +\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1 \qquad (4\text{-}5)$$

(4—5)式表示分布密度曲线下、横轴上的全部面积为1。

第三节　正态分布

正态分布是一种很重要的连续型随机变量的概率分布。生物现象中有许多变量是服从或近似服从正态分布的，如家畜的体长、体重、产奶量、产毛量、血红蛋白含量、血糖含量等。许多统计分析方法都是以正态分布为基础的。此外，还有不少随机变量的概率分布在一定条件下以正态分布为其极限分布。因此在统计学中，正态分布无论在理论研究上还是实际应用中，均占有重要的地位。

一、正态分布的定义及其特征

（一）正态分布的定义　若连续型随机变量x的概率分布密度函数为

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (4\text{-}16)$$

其中μ为平均数，σ^2为方差，则称随机变量x服从正态分布**(normal distribution)**，记为$x \sim N(\mu, \sigma^2)$。相应的概率分布函数为

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \qquad (4\text{-}17)$$

分布密度曲线如图4—2所示。

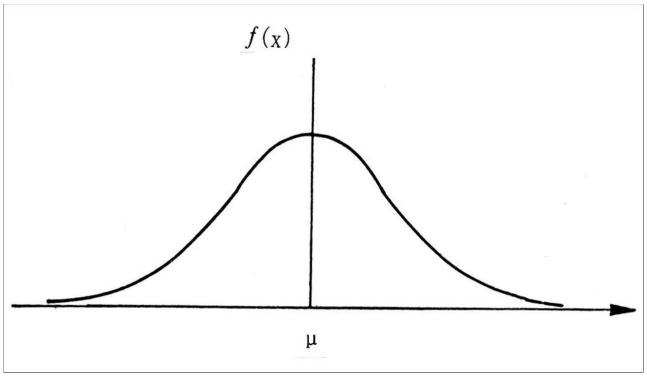


图 4—2　正态分布密度曲线

（二）正态分布的特征 由(4—6)式和图4—2可以看出正态分布具有以下几个重要特征：

1、正态分布密度曲线是单峰、对称的悬钟形曲线，对称轴为$x=\mu$；

2、$f(x)$在$x=\mu$处达到极大，极大值 $f(\mu)=\frac{1}{\sigma\sqrt{2\pi}}$；

3、$f(x)$是非负函数，以x轴为渐近线，分布从$-\infty$至$+\infty$；

4、曲线在$x=\mu\pm\sigma$处各有一个拐点，即曲线在$(-\infty,\mu-\sigma)$和$(\mu+\sigma,+\infty)$区间上是下凸的，在$[\mu-\sigma,\mu+\sigma]$区间内是上凸的；

5、正态分布有两个参数，即平均数μ和标准差σ。μ是位置参数，如图4—3所示。当σ恒定时，μ愈大，则曲线沿x轴愈向右移动；反之，μ愈小，曲线沿x轴愈向左移动。σ是变异度参数，如图4—4所示。当μ恒定时，σ愈大，表示x的取值愈分散，曲线愈“胖”；σ愈小，x的取值愈集中在μ附近，曲线愈“瘦”。

6、分布密度曲线与横轴所夹的面积为1，即：

$$P(-\infty < x < +\infty)=\int_{-\infty}^{+\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx=1$$

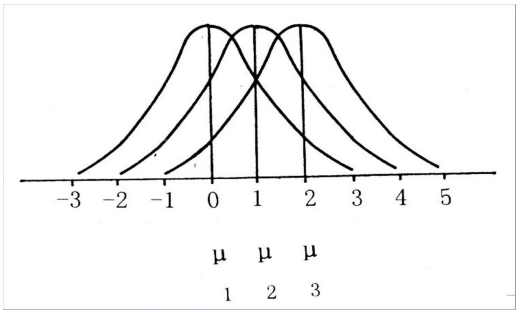


图 4—3 σ相同而μ不同的三个正态分布

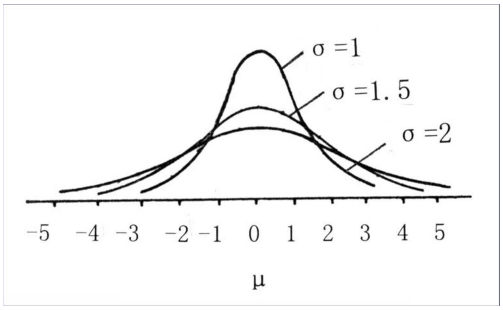


图4—4 μ相同而σ不同的三个正态分

二、标准正态分布

由上述正态分布的特征可知，正态分布是依赖于参数μ和σ^2(或σ)的一簇分布，正态曲线之位置及形态随μ和σ^2的不同而不同。这就给研究具体的正态总体带来困难，需将一般的$N(\mu,\sigma^2)$转换为$\mu=0,\sigma^2=1$的正态分布。我们称$\mu=0,\sigma^2=1$的正态分布为标准正态分布**(standard normal distribution)**。标准正态分布的概率密度函数及分布函数分别记作 $\varphi(u)$和 $\Phi(u)$，由 (4-6)及(4-7) 式得：

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}} \tag{4-8}$$

$$\phi(u)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u}e^{-\frac{1}{2}u^2}du \tag{4-9}$$

随机变量u服从标准正态分布，记作$u\sim N(0,1)$，分布密度曲线如图4—5所示。

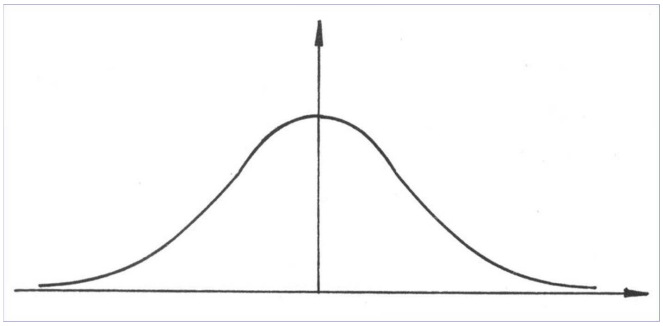


图4—5　　标准正态分布密度曲线

对于任何一个服从正态分布$N(\mu,\sigma^2)$的随机变量x，都可以通过标准化变换：

$$u=(x-\mu)/\sigma \tag{4-10}$$

将其变换为服从标准正态分布的随机变量u。u称为标准正态变量或标准正态离差**(standard normal deviate)**。

按(4-9)式计算，对不同的u值编成函数表，称为正态分布表，见附表1，从中可查到u在意一个区间内取值的概率。这就给解决不同μ、σ²的正态分布概率计算问题带来很大方便。

三、正态分布的概率计算

关于正态分布的概率计算，我们先从标准正态分布着手。这是因为，一方面标准正态分布在正态分布中形式最简单，而且任意正态分布都可化为标准正态分布来计算；另一方面，人们已经根据标准正态分布的分布函数编制成正态分布表(附表1)以供直接查用。

(一) 标准正态分布的概率计算　设u服从标准正态分布，则u在$[u_1,u_2]$内取值的概率为：

$$P(u_1 \le u < u_2)=\frac{1}{\sqrt{2\pi}}\int_{u_1}^{u_2} e^{-\frac{1}{2}u^2}du=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u_2} e^{-\frac{1}{2}u^2}du-\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u_1} e^{-\frac{1}{2}u^2}du$$

$$=\Phi(u_2)-\Phi(u_1) \tag{4-11}$$

而$\Phi(u_1)$与$\Phi(u_2)$可由附表1查得。

附表1只对于$-4.99\le u<4.99$给出了$\Phi(u)$的数值。 表中，u值列在第一列和第一行，第一列列出u的整数部分及小数点后第一位， 第一行为u的小数点后第二位数值 。例如，u=1.75，1.7放在第一列，0.05放在第一行。在附表1中，1.7所在行与0.05 所在列相交处的数值为0.95994，即$\Phi(1.75)=0.95994$。有时会遇到给定$\Phi(u)$值，例如$\Phi(u)=0.284$， 反过来查u值。这只要在附表1中找到与0.284最接近的值0.2843，对应行的第一列数-0.5， 对应列的第一行数值0.07，即相应的u值为$u=-0.57$，亦即$\Phi(-0.57)=0.284$。如果要求更精确的u值，可用线性插值法计算。

表中用了象.$0^3$2336，.$9^3$7674这种写法，分别是0.0002326和0.9997674的缩写，0^3表示连续3个0，9^3表示连续3个9。

由(4-11) 式及正态分布的对称性可推出下列关系式，再借助附表1， 便能很方便地计算有关概率：

$$P(0\leqslant u<u_1)=\Phi(u_1)-0.5$$
$$P(u\geqslant u_1)=\Phi(-u_1)$$
$$P(|u|\geqslant u_1)=2\Phi(-u_1) \qquad (4\text{-}12)$$
$$P(|u|<u_1)=1-2\Phi(-u_1)$$
$$P(u_1\leqslant u<u_2)=\Phi(u_2)-\Phi(u_1)$$

【例4.6】 已知$u\sim N(0,1)$，试求：(1) $P(u<-1.64)=?$ (2) $P(u\geqslant 2.58)=?$ (3) $P(|u|\geqslant 2.56)=?$ (4) $P(0.34\leqslant u<1.53)=?$

利用(4-12)式，查附表1得：

(1) $P(u<-1.64)=0.05050$

(2) $P(u\geqslant 2.58)=\Phi(-2.58)=0.024940$

(3) $P(|u|\geqslant 2.56)=2\Phi(-2.56)=2\times 0.005234=0.010468$

(4) $P(0.34\leqslant u<1.53)=\Phi(1.53)-\Phi(0.34)=0.93669-0.6331=0.30389$

关于标准正态分布，以下几种概率应当熟记：

$$P(-1\leqslant u<1)=0.6826$$
$$P(-2\leqslant u<2)=0.9545$$
$$P(-3\leqslant u<3)=0.9973$$
$$P(-1.96\leqslant u<1.96)=0.95$$
$$P(-2.58\leqslant u<2.58)=0.99$$

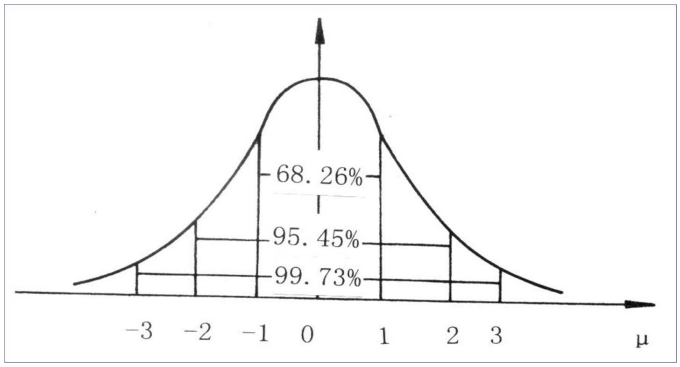


图4—6 标准正态分布的三个常用概率

u变量在上述区间以外取值的概率分别为：

$$P(|u|\geqslant 1)=2\Phi(-1)=1-P(-1\leqslant u<1)=1-0.6826=0.3174$$
$$P(|u|\geqslant 2)=2\Phi(-2)=1-P(-2\leqslant u<2)=1-0.9545=0.0455$$
$$P(|u|\geqslant 3)=1-0.9973=0.0027$$
$$P(|u|\geqslant 1.96)=1-0.95=0.05$$
$$P(|u|\geqslant 2.58)=1-0.99=0.01$$

（二）一般正态分布的概率计算 正态分布密度曲线和横轴围成的一个区域，其面积为1，这实际上表明了“随机变量x取值在$-\infty$与$+\infty$之间”是一个必然事件，其概率为1。若随机变量 x服从正态分布$N(\mu,\sigma^2)$，则x的取值落在任意区间$[x_1,x_2)$的概率，记作$P(x_1\leqslant x<x_2)$，等于图4-7中阴影部分曲边梯形面积。即：

$$P(x_1 \le x < x_2) = \frac{1}{\sigma\sqrt{2\pi}}\int_{x_1}^{x_2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \qquad (4\text{-}13)$$

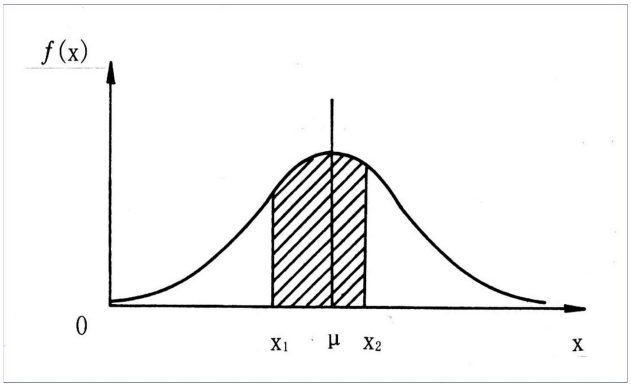


图4—7　正态分布的概率

对 (4-13)式作变换$u=(x-\mu)/\sigma$，得$dx=\sigma\,du$，故有

$$P(x_1 \le u < x_2) = \frac{1}{\sigma\sqrt{2\pi}}\int_{x_1}^{x_2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}du = \frac{1}{\sigma\sqrt{2\pi}}\int_{(x_1-\mu)/\sigma}^{(x_2-\mu)/\sigma} e^{-\frac{1}{2}u^2}\sigma du$$

$$= \frac{1}{\sqrt{2\pi}}\int_{u_1}^{u_2} e^{-\frac{1}{2}u^2}du = \Phi(u_2)-\Phi(u_1)$$

其中，$u_1 = \frac{x_1-\mu}{\sigma}, u_2 = \frac{x_2-\mu}{\sigma}$

这表明服从正态分布$N(\mu,\sigma^2)$的随机变量x在$[x_1, x_2)$内取值的概率，等于服从标准正态分布的随机变量u在$[(x_1-\mu)/\sigma, (x_2-\mu)/\sigma)$内取值的概率。因此，计算一般正态分布的概率时，只要将区间的上下限作适当变换(标准化)，就可用查标准正态分布的概率表的方法求得概率了。

【例4.7】　设x服从$\mu=30.26, \sigma^2=5.10^2$的正态分布，试求$P(21.64\leqslant x<32.98)$。

令$u = \frac{x-30.26}{5.10}$，　则u服从标准正态分布，故

$$P(21.64 \le x < 32.98) = P\left(\frac{21.64-30.26}{5.10} \le \frac{x-30.26}{5.10} < \frac{32.98-30.26}{5.10}\right)$$

$$=P(-1.69\leqslant u<0.53)=\Phi(0.53)-\Phi(-1.69)$$

$$=0.7019-0.04551=0.6564$$

关于一般正态分布，以下几个概率(即随机变量x落在μ加减不同倍数σ区间的概率)是经常用到的。

$$P(\mu-\sigma\leqslant x<\mu+\sigma)=0.6826$$

$$P(\mu-2\sigma\leqslant x<\mu+2\sigma)=0.9545$$

$$P(\mu-3\sigma\leqslant x<\mu+3\sigma)=0.9973$$

$$P(\mu-1.96\sigma\leqslant x<\mu+1.96\sigma)=0.95$$

$$P(\mu-2.58\sigma\leqslant x<\mu+2.58\sigma)=0.99$$

上述关于正态分布的结论，可用一实例来印证。从图2-7可以看出，126头基础母羊体重资料的次数分布接近正态分布，现根据其平均数$\bar{x}=52.26(kg)$，标准差$S=5.10(kg)$，算出平均数加减不同倍数标准差区间内所包括的次数与频率，列于表4—2。

表4—2　126头基础母羊体重在 $\bar{x}\pm kS$ 区间内所包括的次数与频率

$\bar{x}\pm kS$	数 值	区　间	区间内所包含的次数与频率	
			次数	频率（%）
$\bar{x}\pm 1S$	52.26±5.10	47.16－57.36	84	67.46
$\bar{x}\pm 2S$	52.26±10.20	42.06－62.46	119	94.44
$\bar{x}\pm 3S$	52.26±15.30	36.96－67.56	126	100.00
$\bar{x}\pm 1.96S$	52.26±10.00	42.26－62.26	119	94.44
$\bar{x}\pm 2.58S$	52.26±13.16	39.10－65.42	126	100.00

由表4—2可见，实际频率与理论概率相当接近，说明126头基础母羊体重资料的频率分布接近正态分布，从而可推断基础母羊体重这一随机变量很可能是服从正态分布的。

生物统计中，不仅注意随机变量x落在平均数加减不同倍数标准差区间(μ－kσ，μ＋kσ)之内的概率而且也很关心x落在此区间之外的概率。我们把随机变量x落在平均数μ加减不同倍数标准差σ区间之外的概率称为双侧概率(两尾概率)，记作α。对应于双侧概率可以求得随机变量x小于μ－kσ或大于μ＋kσ的概率，称为单侧概率(一尾概率),记作α／2。例如，x落在(μ－1.96σ，μ＋1.96σ)之外的双侧概率为0.05，而单侧概率为0.025。即

$$P(x<\mu-1.96\sigma)=P(x>\mu+1.96\sigma)=0.025$$

双侧概率或单侧概率如图4—8所示。x落在(μ－2.58σ，μ＋2.58σ)之外的双侧概率为0.01，而单侧概率

$$P(x<\mu-2.58\sigma)=P(x>\mu+2.58\sigma)=0.005$$

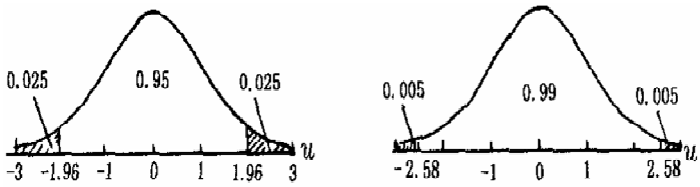


图4—8　双侧概率与单侧概率

附表2给出了满足P（｜u｜＞u_α)=α的双侧分位u_α的数值。因此，　只要已知双侧概率α的值，由附表2就可直接查出对应的双侧分位数u_α，查法与附表1相同。例如，已知u～$N(0,1)$试求：

(1)　$P(u<-u_\alpha)+P(u\geq u_\alpha)=0.10$的$u_\alpha$

(2)　$P(-u_\alpha\leq u<u_\alpha)=0.86$的$u_\alpha$

因为附表2中的α值是：

$$\alpha=1-\frac{1}{\sqrt{2\pi}}\int_{-u_\alpha}^{u_\alpha}e^{-\frac{1}{2}u^2}du$$

所以

(1) $P(u<-u_\alpha)+P(u\geq u_\alpha)=1-P(-u_\alpha\leq u<u_\alpha)=0.10=\alpha$

由附表2查得：$u_{0.10}=1.644854$

(2)　$P(-u_\alpha\leq u<u_\alpha)=0.86$，　$\alpha=1-P(-u_\alpha\leq u<u_\alpha)=1-0.86=0.14$

由附表2查得：$u_{0.14}$=1.475791

对于$x\sim N(\mu,\sigma^2)$，只要将其转换为$u\sim N(0,1)$,即可求得相应的双侧分位数。

【例4.8】 已知猪血红蛋白含量x服从正态分布N(12.86，1.33^2)， 若$P(x<l_1)$=0.03, $P(x\geqslant l_2)$=0.03,求l_1，l_2。

由题意可知，α／2=0.03，α=0.06 又因为

$$P(x<l_1)=P(\frac{x-12.86}{1.33}<\frac{l_1-12.86}{1.33})=P(u<-u_\alpha)=0.03$$

$$P(x\geqslant l_2)=P(\frac{x-12.86}{1.33}\geq\frac{l_2-12.86}{1.33})=P(u\geq u_\alpha)=0.03$$

故 $P(x<l_1=+P(x\geqslant l_2)=P(u<-u_\alpha=+P(u\geqslant u_\alpha)$

$=1-P(-u_\alpha\leqslant P<u_\alpha)=0.06=\alpha$

由附表2查得：$u_{0.06}$=1.880794,所以

$(l_1$-12.86)/1.33=-1.880794， $(l_2$-12.86)/1.33=1.880794

即 $l_1\approx$10.36， $l_2\approx$15.36。

第四节 二项分布

一、贝努利试验及其概率公式

将某随机试验重复进行n次，若各次试验结果互不影响， 即每次试验结果出现的概率都不依赖于其它各次试验的结果，则称这n次试验是独立的。

对于n次独立的试验，如果每次试验结果出现且只出现对立事件A与$\overline{A}$之一，在每次试验中出现A的概率是常数$p(0<p<1)$,因而出现对立事件$\overline{A}$的概率是$1-p=q$，则称这一串重复的独立试验为n重贝努利试验，简称贝努利试验**(Bernoulli trials)**。

在生物学研究中，我们经常碰到的一类离散型随机变量，如入孵n枚种蛋的出雏数、n头病畜治疗后的治愈数、n 尾鱼苗的成活数等，可用贝努利试验来概括。

在n重贝努利试验中，事件A可能发生0，1，2，…，n次，现在我们来求事件A恰好发生$k(0\leqslant k\leqslant n)$次的概率$P_n(k)$。

先取n=4，k=2来讨论。在4次试验中，事件A发生2次的方式有以下C_4^2种：

$A_1A_2\,\overline{A}_3\overline{A}_4$　　$A_1\overline{A}_2A_3\overline{A}_4$　　$A_1\overline{A}_2\overline{A}_3A_4$

$\overline{A}_1A_2A_3\overline{A}_4$　　$\overline{A}_1A_2\overline{A}_3A_4$　　$\overline{A}_1\overline{A}_2A_3A_4$

其中$A_k(k=1,2,3,4)$表示事件A在第k次试验发生；$\overline{A}_k$ (k=1, 2, 3, 4)表示事件A在第k次试验不发生。由于试验是独立的，按概率的乘法法则，于是有

$$P(A_1A_2\,\overline{A}_3\overline{A}_4)=P(A_1\overline{A}_2A_3\overline{A}_4)=\cdots=P(\overline{A}_1\overline{A}_2A_3A_4)$$

$$=P(A_1)\cdot P(A_2)\cdot P(\overline{A}_3)\cdot P(\overline{A}_4)=p^2q^{4-2}$$

又由于以上各种方式中，任何二种方式都是互不相容的，按概率的加法法则，在4 次试验中，事件A恰好发生2次的概率为

$$P_4(2)=P(A_1A_2\,\overline{A}_3\overline{A}_4)+P(A_1\overline{A}_2A_3\overline{A}_4)+\cdots+P(\overline{A}_1\overline{A}_2A_3A_4)=C_4^2\;p^2q^{4-2}$$

一般，在n重贝努利试验中，事件A恰好发生k(0≤k≤n)次的概率为

$$P_n(k)=C_n^k\ p^k q^{n-k} \qquad k=0,1,2\cdots,\ n \tag{4-14}$$

若把(4-14)式与二项展开式

$$(q+p)^n=\sum_{k=0}^{n} C_n^k\ p^k q^{n-k}$$

相比较就可以发现，在n重贝努利试验中，事件A发生k次的概率恰好等于$(q+p)^n$ 展开式中的第k+1项，所以也把(4-14)式称作二项概率公式。

二、二项分布的意义及性质

二项分布定义如下：

设随机变量x所有可能取的值为零和正整数：0,1,2,…，n，且有

$$P_n(k)=C_n^k\ p^k q^{n-k} \qquad k=0,1,2\cdots,\ n$$

其中$p>0$，$q>0$，$p+q=1$，则称随机变量x服从参数为n和p的二项分布 **(binomial distribution)**，记为 $x\sim B(n,p)$。

显然，二项分布是一种离散型随机变量的概率分布。参数n称为离散参数， 只能取正整数；p是连续参数，它能取0与1之间的任何数值(q由p确定，故不是另一个独立参数)。

容易验证，二项分布具有概率分布的一切性质，即：

1、$P(x=k)=P_n(k)\ \geq 0$ （$k=0,1,\ldots,\ n$）

2、二项分布的概率之和等于1，即

$$\sum_{k=0}^{n} C_n^k\ p^k q^{n-k}=(q+p)^n=1$$

3、$$P(x\leq m)=P_n(k\leq m)=\sum_{k=0}^{m} C_n^k\ p^k q^{n-k} \tag{4-15}$$

4、$$P(x\geq m)=P_n(k\geq m)=\sum_{k=m}^{n} C_n^k\ p^k q^{n-k} \tag{4-16}$$

5、$$P(m_1\leq x\leq m_2)=p_n(m_1\leq k\leq m_2)=\sum_{k=m_1}^{m_2} C_n^k\ p^k q^{n-k} \quad (m_1<m_2) \tag{4-17}$$

二项分布由n和p两个参数决定：

1、当p值较小且n不大时，分布是偏倚的。但随着n的增大 ，分布逐渐趋于对称，如图4—9 所示；

2、当p值趋于0.5时，分布趋于对称，如图4—10所示；

3、对于固定的n及p，当k增加时，$P_n(k)$先随之增加并达到其极大值，以后又下降。

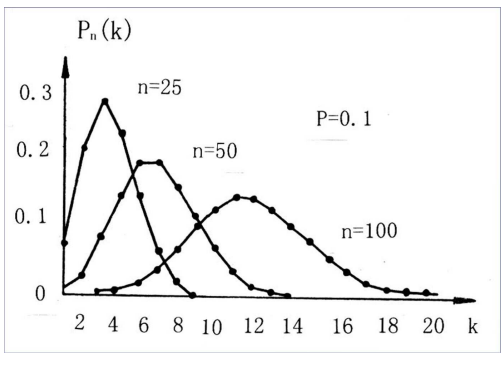


图 4—9 n值不同的二项分布比较

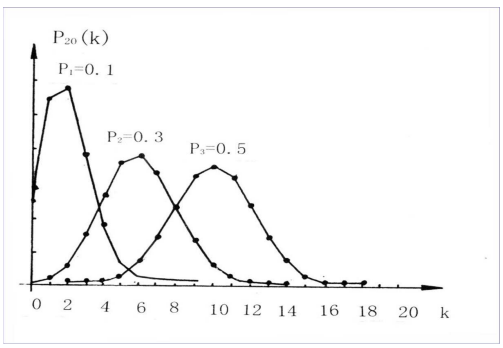


图4—10 n值不同的二项分布比较

此外，在n较大，np、nq较接近时，二项分布接近于正态分布；当$n\to\infty$时，二项分布的极限分布是正态分布。

三、二项分布的概率计算及应用条件

【例4.9】 纯种白猪与纯种黑猪杂交，根据孟德尔遗传理论， 子二代中白猪与黑猪的比率为3∶1。求窝产仔10头，有7头白猪的概率。

根据题意，n=10，p=3 / 4=0.75，q=1 / 4=0.25。设10头仔猪中白色的为x头，则x为服从二项分布B(10，0.75)的随机变量。于是窝产10头仔猪中有7头是白色的概率为：

$$P(x=7)=C_{10}^{7}0.75^{7}0.25^{3}=\frac{10!}{7!3!}\times0.75^{7}\times0.25^{3}=0.2503$$

【例4.10】 设在家畜中感染某种疾病的概率为20%，现有两种疫苗，用疫苗A注射了15头家畜后无一感染，用疫苗B注射15头家畜后有1头感染。设各头家畜没有相互传染疾病的可能，问：应该如何评价这两种疫苗?

假设疫苗A完全无效，那么注射后的家畜感染的概率仍为20%，则15 头家畜中染病头数x=0的概率为

$$p(x=0)=C_{15}^{0}0.20^{0}0.80^{15}=0.0352$$

同理，如果疫苗B完全无效，则15头家畜中最多有1头感染的概率为

$$p(x\le1)=C_{15}^{0}0.2^{0}0.8^{15}+C_{15}^{1}0.2^{1}0.8^{14}=0.1671$$

由计算可知，注射A疫苗无效的概率为0.0352，比B疫苗无效的概率0.1671小得多。因此，可以认为A疫苗是有效的，但不能认为B疫苗也是有效的。

【例4.11】 仔猪黄痢病在常规治疗下死亡率为20%，求5 头病猪治疗后死亡头数各可能值相应的概率。

设5头病猪中死亡头数为x，则x服从二项分布B(5,0.2)，其所有可能取值为0，1，…，5,按(4-6)式计算概率用分布列表示如下：

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 0.3277 & 0.4096 & 0.2048 & 0.0512 & 0.0064 & 0.0003 \end{pmatrix}$$

从上面各例可看出二项分布的应用条件有三：（1）各观察单位只具有互相对立的一种结果，如阳性或阴性，生存或死亡等，属于二项分类资料；（2）已知发生某一结果(如死亡）的概率为p，其对立结果的概率则为1-P=q，实际中要求p 是从大量观察中获得的比较稳定的数值；（3）n个观察单位的观察结果互相独立，即每个观察单位的观察结果不会影响到其它观察单位的观察结果。

四、二项分布的平均数与标准差

前面已经指出二项分布由两个参数n和p决定。统计学证明，服从二项分布$B(n,p)$的随机变量之平均数μ、标准差σ与参数n、p有如下关系：

当试验结果以事件A发生次数k表示时

$$\mu = np \tag{4-18}$$

$$\sigma = \sqrt{npq} \tag{4-19}$$

【例4.12】 求**【例4.11】**平均死亡猪数及死亡数的标准差。

以p=0.2,n=5代入 (4-18)和(4-19) 式得

平均死亡猪数 μ=5×0.20=1.0(头)

标准差 $\sigma = \sqrt{npq} = \sqrt{5\times0.2\times0.8} = 0.894$(头)

当试验结果以事件A发生的频率k/n表示时

$$\mu_p = p \tag{4-20}$$

$$\sigma_p = \sqrt{(pq)/n} \tag{4-21}$$

σ_p也称为总体百分数标准误，当p未知时，常以样本百分数$\hat{p}$来估计。此时(4-21) 式改写为：

$$S_p = \sqrt{(\hat{p}\hat{q})/n} \qquad \hat{q} = 1 - \hat{p} \tag{4-22}$$

S_p称为样本百分数标准误。

第五节 波松分布

波松分布是一种可以用来描述和分析随机地发生在单位空间或时间里的稀有事件的概率分布。要观察到这类事件，样本含量n必须很大 。在生物、医学研究中，服从波松分布的随机变量是常见的。如，一定畜群中某种患病率很低的非传染性疾病患病数或死亡数，畜群中遗传的畸形怪胎数，每升饮水中大肠杆菌数，计数器小方格中血球数，单位空间中某些野生动物或昆虫数，医院门诊单位时间内就诊患者数等，都是服从波松分布的。

一、波松分布的意义

若随机变量$x(x=k)$只取零和正整数值0，1，2，…，且其概率分布为

$$P(x=k) = \frac{\lambda^k}{k!}e^{-\lambda} \quad ，k=0，1，\cdots\cdots \tag{4-23}$$

其中λ＞0；e=2.7182…是自然对数的底数，则称x服从参数为λ的波松分布**(Poisson's distribution)**，记为$x\sim P(\lambda)$。

波松分布作为一种离散型随机变量的概率分布有一个重要的特征，这就是它的平均数和方差相等，都等于常数λ，即$\mu=\sigma^2=\lambda$。利用这一特征， 可以初步判断一个离散型随机变量是否服从波松分布。

【例4.13】 调查某种猪场闭锁育种群仔猪畸形数，共记录200窝， 畸形仔猪数的分布情况如表4-3所示。试判断畸形仔猪数是否服从波松分布。

表4-3　畸形仔猪数统计分布

每窝畸形数k	0	1	3	3	≥4	合计
窝　数 f	120	62	15	2	1	200

根据波松分布的平均数与方差相等这一特征，若畸形仔猪数服从波松分布，则由观察数据计算的平均数和方差就近于相等。样本均数 $\bar{x}$ 和方差S^2计算结果如下：

$\bar{x}=\Sigma fk/n=(120\times0+62\times1+15\times2+2\times3+1\times4)/200=0.51$-

$$s^2=\frac{\sum fk^2-(\sum fk)^2/n}{n-1}=\frac{(120\times0^2+62\times1^2+15\times2^2+2\times3^2+1\times4^2-102^2)/200}{200-1}=0.52$$

$\bar{x}$ =0.51，S^2=0.52,这两个数是相当接近的， 因此可以认为畸形仔猪数服从波松分布。

λ是波松分布所依赖的唯一参数。λ值愈小分布愈偏倚，随着λ的增大，分布趋于对称(如图4-11所示)。当λ=20时分布接近于正态分布；当λ=50时，可以认为波松分布呈正态分布。所以在实际工作中，当λ≥20时就可以用正态分布来近似地处理波松分布的问题。

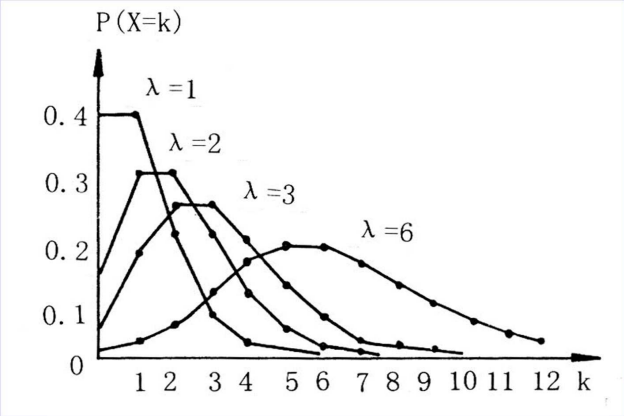


图4—11　不同λ的波松分布

二、波松分布的概率计算

由(4-23)式可知，波松分布的概率计算，依赖于参数λ的确定，只要参数λ确定了，把k=0,1,2,…代入(4-23)式即可求得各项的概率。 但是在大多数服从波松分布的实例中，分布参数λ往往是未知的，只能从所观察的随机样本中计算出相应的样本平均数作为λ的估计值，将其代替(4-23)式中的λ，计算出k=0,1,2,…时的各项概率。

如【例4.13】中已判断畸形仔猪数服从波松分布，并已算出样本平均数 $\bar{x}$ =0.51。将0.51代替公式（4-23）中的λ得：

$$P(x=k)=\frac{0.51^k}{k!}e^{-0.51}\qquad(k=0,1,2,\cdots)$$

因为$e^{-0.51}$=1.6653，所以畸形仔猪数各项的概率为：

$P(x=0)=0.51^0/(0!\times1.6653)=0.6005$

$P(x=1)=0.51^1/(1!\times1.6653)=0.3063$

$P(x=2)=0.51^2/(2!\times1.6653)=0.0781$

$P(x=3)=0.51^3/(3!\times1.6653)=0.0133$

$P(x=4)=0.51^4/(4!\times1.6653)=0.0017$

$$P(x>4)=1-\sum_{k=0}^{4}p(x=k)=1-0.9999=0.0001$$

把上面各项概率乘以总观察窝数(N=200)即得各项按波松分布的理论窝数。波松分布与相应的频率分布列于表4—7中。

表4—4　畸形仔猪数的波松分布

每窝畸形数 k	0	1	2	3	≥4	合计
窝　数	120	62	15	2	1	200
频　率	0.6000	0.3100	0.0750	0.0100	0.0050	1.00
概　率	0.6005	0.3063	0.0781	0.0133	0.0018	1.00
理论窝数	120.12	61.26	15.62	2.66	0.34	200

将实际计算得的频率与根据λ=0.51的泊松分布计算的概率相比较，发现畸形仔猪的频率分布与λ=0.51的波松分布是吻合得很好的。这进一步说明了畸形仔猪数是服从波松分布的。

【例4.14】　为监测饮用水的污染情况，现检验某社区每毫升饮用水中细菌数，共得400个记录如下：

1ml水中细菌数	0	1	2	≥3	合　计
次数f	243	120	31	6	400

试分析饮用水中细菌数的分布是否服从波松分布。若服从，按波松分布计算每毫升水中细菌数的概率及理论次数并将次数分布与波松分布作直观比较。

经计算得每毫升水中平均细菌数$\bar{x}$=0.500,方差S^2=0.496。两者很接近，故可认为每毫升水中细菌数服从波松分布。以$\bar{x}$=0.500代替（4-23）式中的λ，得

$$P(x=k)=\frac{0.5^{k}}{k!}e^{-0.5}\qquad(k=0,1,2\cdots)$$

计算结果如表4—5所示。

表4—5　细菌数的波松分布

1ml水中细菌数	0	1	2	≥3	合　计
实际次数	243	120	31	6	400
频　率	0.6075	0.3000	0.0775	0.0150	1.00
概　率	0.6065	0.3033	0.0758	0.0144	1.00
理论次数	242.60	121.32	30.32	5.76	400

可见细菌数的频率分布与λ=0.5的波松分布是相当吻合的，进一步说明用波松分布描述单位容积(或面积)中细菌数的分布是适宜的。

应当注意，二项分布的应用条件也是波松分布的应用条件。比如二项分布要求n 次试验是相互独立的，这也是波松分布的要求。然而一些具有传染性的罕见疾病的发病数，因为首例发生之后可成为传染源，会影响到后续病例的发生，所以不符合波松分布的应用条件。对于在单位时间、单位面积或单位容积内，所观察的事物由于某些原因分布不随机时，如细菌在牛奶中成集落存在时，亦不呈波松分布。

前面讨论的三个重要的概率分布中，前一个属连续型随机变量的概率分布，后两个属

离散型随机变量的概率分布。三者间的关系如下：

对于二项分布，在$n\to\infty, p\to 0$，且$np=\lambda$（较小常数）情况下，二项分布趋于波松布。在这种场合，波松分布中的参数λ用二项分布的np代之；在$n\to\infty$，$p\to 0.5$时，二项分布趋于正态分布。在这种场合，正态分布中的μ、σ^2用二项分布的np、npq代之。在实际计算中，当$p<0.1$且n很大时，二项分布可由波松分布近似；当$p>0.1$且n很大时，二项分布可由正态分布近似。

对于波松分布，当$\lambda\to\infty$时，波松分布以正态分布为极限。在实际计算中，当$\lambda\geqslant 20$（也有人认为$\lambda\geqslant 6$）时，用波松分布中的λ代替正态分布中的μ及σ^2，即可由后者对前者进行近似计算。

第六节　样本平均数的抽样分布

研究总体与从中抽取的样本之间的关系是统计学的中心内容。对这种关系的研究可从两方面着手，一是从总体到样本，这就是研究抽样分布**(sampling distribution)**的问题；二是从样本到总体，这就是统计推断**(statistical inference)**问题。 统计推断是以总体分布和样本抽样分布的理论关系为基础的。为了能正确地利用样本去推断总体，并能正确地理解统计推断的结论，须对样本的抽样分布有所了解。

我们知道，由总体中随机地抽取若干个体组成样本，即使每次抽取的样本含量相等，其统计量（如$\bar{x}$，S）也将随样本的不同而有所不同，因而样本统计量也是随机变量， 也有其概率分布。我们把统计量的概率分布称为抽样分布。本节仅就样本平均数的抽样分布加以讨论。

一、样本平均数抽样分布

由总体随机抽样**(random sampling)**的方法可分为有返置抽样和不返置抽样两种。 前者指每次抽出一个个体后，这个个体应返置回原总体；后者指每次抽出的个体不返置回原总体。对于无限总体，返置与否都可保证各个体被抽到的机会相等。对于有限总体，就应该采取返置抽样，否则各个体被抽到的机会就不相等。

设有一个总体，总体平均数为μ，方差为σ^2，总体中各变数为x， 将此总体称为原总体。现从这个总体中随机抽取含量为n的样本，样本平均数记为$\bar{x}$。可以设想，从原总体中可抽出很多甚至无穷多个含量为n的样本。由这些样本算得的平均数有大有小，不尽相同，与原总体平均数μ相比往往表现出不同程度的差异。这种差异是由随机抽样造成的， 称为抽样误差**(sampling error)**。显然，样本平均数也是一个随机变量，其概率分布叫做样本平均数的抽样分布。由样本平均数$\bar{x}$构成的总体称为样本平均数的抽样总体，其平均数和标准差分别记为$\mu_{\bar{x}}$和$\sigma_{\bar{x}}$。$\sigma_{\bar{x}}$是样本平均数抽样总体的标准差，简称标准误**(standard error)**，它表示平均数抽样误差的大小。统计学上已证明$\bar{x}$总体的两个参数与x 总体的两个参数有如下关系：

$$\mu_{\bar{x}}=\mu,\quad \sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}} \tag{4—24}$$

为了验证这个结论及了解平均数抽样总体与原总体概率分布间的关系，我们进行一个模拟抽样试验。

设有一个N=4的有限总体，变数为2、3、3、4。根据$\mu=\Sigma x/N$和$\sigma^2=\Sigma(x-\mu)^2/N$求得该总体的$\mu$、$\sigma^2$、$\sigma$为：

$$\mu=3,\quad \sigma^2=1/2,\quad \sigma=\sqrt{1/2}=0.707$$

从有限总体作返置随机抽样，所有可能的样本数为N^n个，其中n为样本含量。以上述总体而论，如果从中抽取n=2的样本，共可得4^2=16个样本；如果样本含量n为4，则一共可抽得4^4=256个样本。分别求这些样本的平均数$\bar{x}$，其次数分布如表4—6所示。

根据表4—6，在n=2的试验中，样本平均数抽样总体的平均数、方差与标准差分别为：

$$\mu_{\bar{x}}=\sum f\bar{x}/N^n=48.0/16=3=\mu$$

$$\sigma_{\bar{x}}^2=\frac{\sum f(\bar{x}-\mu_{\bar{x}})^2}{N^n}=\frac{\sum f\bar{x}^2-(\sum f\bar{x})^2/N^n}{N^n}=\frac{148-48^2/16}{16}$$

$$=4/16=1/4=(1/2)/2=\sigma^2/n$$

$$\sigma_{\bar{x}}=\sqrt{\sigma_{\bar{x}}^2}=\sqrt{1/4}=\sqrt{1/2}/\sqrt{2}=\sigma/\sqrt{n}$$

表4—6　N=4, n=2和n=4时$\bar{x}$的次数分布

$N^n=4^2=16$				$N^n=4^4=256$			
$\bar{x}$	f	$f\bar{x}$	$f\bar{x}^2$	$\bar{x}$	f	$f\bar{x}$	$f\bar{x}^2$
2.0	1	2.0	4.00	2.00	1	2.00	4.0000
2.5	4	10.0	25.00	2.25	8	18.00	40.5000
3.0	6	18.0	54.00	2.50	28	70.00	175.0000
3.5	4	14.0	49.00	2.75	56	154.00	423.5000
4.0	1	4.0	16.00	3.00	70	210.00	630.0000
				3.25	56	182.00	591.5000
				3.50	28	98.00	343.0000
				3.75	8	30.00	112.5000
				4.00	1	4.00	16.0000
Σ	16	48.0	148.00	Σ	256	768.00	2336.0000

同理，可得n=4时：

$$\mu_{\bar{x}}=768/256=3=\mu \qquad \sigma_{\bar{x}}^2=32/256=1/8=(1/2)/4=\sigma^2/n$$

$$\sigma_{\bar{x}}=\sqrt{1/8}=\sqrt{1/2}\Big/\sqrt{4}=\sigma/\sqrt{n}$$

这就验证了$\mu_{\bar{x}}=\mu$，$\sigma_{\bar{x}}=\sigma/\sqrt{n}$的正确性。

若将表4—6中两个样本平均数的抽样总体作次数分布图，则如图4-12所示。

由以上模拟抽样试验可以看出，虽然原总体并非正态分布，但从中随机抽取样本，即使样本含量很小(n=2, n=4)，样本平均数的分布却趋向于正态分布形式。随着样本含量n的增大，样本平均数的分布愈来愈从不连续趋向于连续的正态分布。比较图4—12两个分

布，在n由2增到4时，这种趋势表现得相当明显。当$n>30$时，$\bar{x}$ 的分布就近似正态分布了。x变量与 $\bar{x}$ 变量概率分布间的关系可由下列两个定理说明：

1. 若随机变量x服从正态分布$N(\mu, \sigma^2)$，$x_1, x_2, \cdots, x_n$ 是由x总体得来的随机样本，则统计量 $\bar{x}=\Sigma x/n$的概率分布也是正态分布， 且有 $\mu_{\bar{x}}=\mu$， $\sigma_{\bar{x}}=\sigma/\sqrt{n}$， 即 $\bar{x}$ 服从正态分布$N(\mu, \sigma^2/n)$。

2. 若随机变量x服从平均数是μ，方差是σ^2的分布(不是正态分布)； $x_1, x_2, \cdots, x_n$ 是由此总体得来的随机样本，则统计量 $\bar{x}=\Sigma x/n$的概率分布，当n相当大时逼近正态分布 $N(\mu, \sigma^2/n)$。这就是中心极限定理。

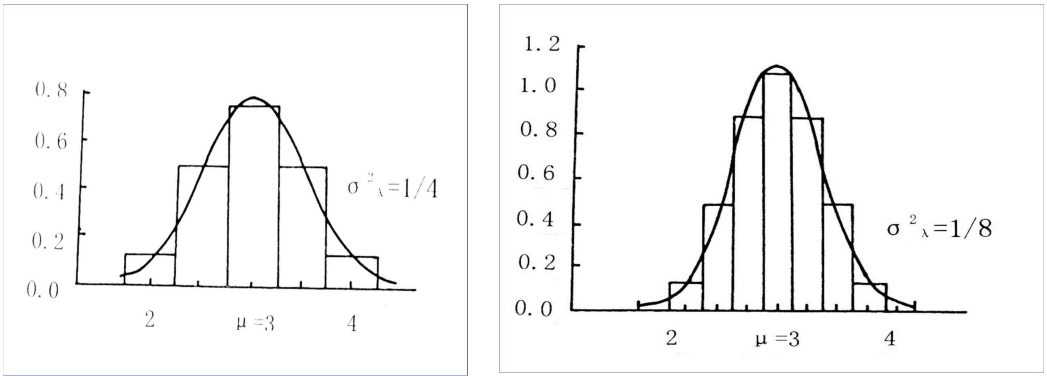


图 4-12　平均数 $\bar{x}$ 的抽样分布

上述两个结果保证了样本平均数的抽样分布服从或者逼近正态分布。

中心极限定理告诉我们：不论x变量是连续型还是离散型，也无论x服从何种分布，一般只要$n>30$，就可认为 $\bar{x}$ 的分布是正态的。若x的分布不很偏倚，在$n>20$时， $\bar{x}$ 的分布就近似于正态分布了。这就是为什么正态分布较之其它分布应用更为广泛的原因。

二、标　准　误

标准误(平均数抽样总体的标准差) $\sigma_{\bar{x}}=\sigma/\sqrt{n}$ 的大小反映样本平均数 $\bar{x}$ 的抽样误差的大小，即精确性的高低。标准误大，说明各样本平均数 $\bar{x}$ 间差异程度大，样本平均数的精确性低。反之，$\sigma_{\bar{x}}$ 小，说明 $\bar{x}$ 间的差异程度小，样本平均数的精确性高。$\sigma_{\bar{x}}$ 的大小与原总体的标准差σ成正比，与样本含量n的平方根成反比。从某特定总体抽样，因为σ是一常数，所以只有增大样本含量才能降低样本平均数 $\bar{x}$ 的抽样误差。

在实际工作中，总体标准差σ往往是未知的，因而无法求得$\sigma_{\bar{x}}$。此时，可用样本标准差S估计σ。于是，以 $S/\sqrt{n}$ 估计 $\sigma_{\bar{x}}$。记 $S/\sqrt{n}$ 为 $S_{\bar{x}}$，称作样本标准误或均数标准误。样本标准误 $S_{\bar{x}}$ 是平均数抽样误差的估计值。若样本中各观测值为 $x_1, x_2, \cdots, x_n$，则

$$S_{\bar{x}}=\frac{S}{\sqrt{n}}=\sqrt{\frac{\sum(x-\bar{x})^2}{n(n-1)}}=\sqrt{\frac{\sum x^2-(\sum x)^2/n}{n(n-1)}} \qquad (4\text{-}25)$$

应当注意，样本标准差与样本标准误是既有联系又有区别的两个统计量，(4—25) 式已表明了二者的联系。二者的区别在于：样本标准差S是反映样本中各观测值 $x_1, x_2, \cdots, x_n$ 变异程度大小的一个指标，它的大小说明了 $\bar{x}$ 对该样本代表性的强弱。样本标准误是样

本平均数 $\bar{x}_1,\bar{x}_2,...,\bar{x}_k$ 的标准差，它是 $\bar{x}$ 抽样误差的估计值， 其大小说明了样本间变异程度的大小及 $\bar{x}$ 精确性的高低。

对于大样本资料，常将样本标准差S与样本平均数 $\bar{x}$ 配合使用，记为 $\bar{x}\pm S$，用以说明所考察性状或指标的优良性与稳定性。对于小样本资料，常将样本标准误 $S_{\bar{x}}$ 与样本平均数 $\bar{x}$ 配合使用，记为 $\bar{x}\pm S_{\bar{x}}$，用以表示所考察性状或指标的优良性与抽样误差的大小。

第七节 t 分 布

由样本平均数抽样分布的性质知道： 若x～N(μ，σ²)， 则 $\bar{x}$ ～$N(\mu,\ \sigma^2/n)$。 将随机变量 $\bar{x}$ 标准化得：$u=(\bar{x}-\mu)/\sigma_{\bar{x}}$，则$u$～$N(0,1)$。 当总体标准差σ未知时， 以样本标准差$S$代替σ所得到的统计量 $(\bar{x}-\mu)/S_{\bar{x}}$ 记为t。在计算 $S_{\bar{x}}$ 时，由于采用S来代替σ，使得 t 变量不再服从标准正态分布，而是服从t分布 **(t－distribution)**。它的概率分布密度函数如下：

$$f(t)=\frac{1}{\sqrt{\pi df}}\frac{\Gamma[(df+1)/2]}{\Gamma(df/2)}(1+\frac{t^2}{df})^{-\frac{df+1}{2}} \qquad (4\text{-}26)$$

式中，t的取值范围是（-∞，+∞）；df=n-1为自由度。

t分布的平均数和标准差为：

$$\mu_t=0\ (df>1),\quad \sigma_t=\sqrt{df/(df-2)}\ (df>2) \qquad (4\text{-}27)$$

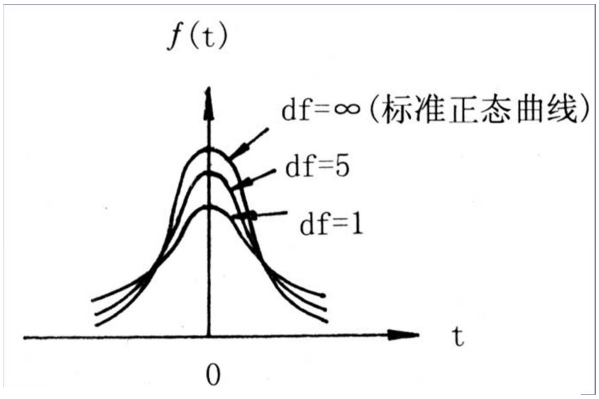


图4-13 不同自由度的t分布密度曲线

t分布密度曲线如图4-13所示，其特点是：

1、t分布受自由度的制约，每一个自由度都有一条t分布密度曲线。

2、t分布密度曲线以纵轴为对称轴，左右对称，且在t=0时，分布密度函数取得最大值。

3、与标准正态分布曲线相比，t分布曲线顶部略低，两尾部稍高而平。df越小这种趋势越明显。df越大，t分布越趋近于标准正态分布。当n >30时，t分布与标准正态分布的区别很小；n >100时，t分布基本与标准正态分布相同；$n\to\infty$时，t 分布与标准正态分布完全一致。

t分布的概率分布函数为：

$$F_{t(df)} = P(t < t_1) = \int_{-\infty}^{t_1} f(t)dt \qquad (4\text{-}28)$$

因而t在区间（t_1，+∞）取值的概率——右尾概率为$1-F_t(df)$。由于t分布左右对称，t在区间（-∞，$-t_1$）取值的概率也为$1-F_t(df)$。于是t分布曲线下由-∞到$-t_1$和由t_1到+∞两个相等的概率之和——两尾概率为$2(1-F_t(df))$。对于不同自由度下t分布的两尾概率及其对应的临界t值已编制成附表3，即t分布表。该表第一列为自由度df，表头为两尾概率值，表中数字即为临界t值。

例如，当df=15时，查附表3得两尾概率等于0.05的临界t值为$t_{0.05(15)}$ =2.131，其意义是：$P(-\infty<t<-2.131)=P(2.131<t<+\infty)=0.025$；$P(-\infty<t<-2.131)+P(2.131<t<+\infty)=0.05$。

由附表3可知，当df一定时，概率P越大，临界t值越小；概率P越小，临界t值越大。 当概率P一定时，随着df的增加，临界t值在减小，当df=∞时，临界t值与标准正态分布的临界u值相等。

习　　题

1、什么是随机试验？它具有那三个特征？

2、什么是必然事件、不可能事件、随机事件?

3、概率的统计定义及古典定义分别是什么？事件的概率具有那些基本性质？

4、什么是小概率事件实际不可能性原理？

5、袋中有10只乒乓球，分别编浩为1到10，从中随机抽取3只记录其编号。

(1) 求最小的号码为5的概率；(1/12)

(2) 求最大的号码为5的概率；(1/20)

6、现有6只雏鸡，其中4只是雌的，2只是雄的，从中抽取两次，每次取一只，在返回抽样情况下求：

（1）取到的两只雏鸡都是雌性的概率；

（2）取到的两只雏鸡性别相同的概率；

（3）取到的两只雏鸡至少有一只是雌性的概率；

[（1）0.444；　（2）0.556；　（3）0.889]

7、假设每个人的血清中含有肝炎病毒的概率为0.4%，混和100个人的血清， 求此血清中含有肝炎病毒的概率。你认为计算结果会告诉我们一个什么事实?(0.33)

8、离散型随机变量概率分布与连续型随机变量概率分布有何区别？

9、什么是正态分布?标准正态分布?正态分布的密度曲线有何特点?

10、已知随机变量u服从N(0，1)，求P(u<-1.4)，　P(u≥1.49)，　P（|u|≥2.58），　P(-1.21≤u<0.45)，并作图示意。(0.0792，0.06811，0.00988，0.5605)

11、已知随机变量u服从N(0，1)，求下列各式的u_α。

(1) P(u<$-u_\alpha$)+P(u≥u_α)=0.1;0.52

(2) P($-u_\alpha$≤u<u_α)=0.42;0.95

[（1）1.644854,0.643345；（2）0.553385,1.959964]]

12、猪血红蛋白含量x服从正态分布N(12.86，1.33^2)

(1) 求猪血红蛋白含量x在11.53—14.19范围内的概率。

(2) 若P(x<l_1)=0.025，P(x>l_2)=0.025，求l_1，l_2。

[（1）0.6826，（2）l_1=10.25，l_2=15.47]

13、设x变量服从正态分布，总体平均数μ=10,P(x≥12)=0.1056,试求x在区间6—16内取值的概率。(0.914948)

14、什么是二项分布?如何计算二项分布的平均数、方差和标准差?

15、已知随机变量x服从二项分布B（100，0.1），求μ及σ。(10，3)

16、记录表明，10头家畜已有3头死于某种疾病，现有5头病畜，试求以下情况的概率：

⑴ 恰有3头死亡；　　(1323/10000)

⑵ 前面3头死亡，后2头康复；　　(1323/100000)

⑶ 前面3头死亡；　　(27/1000)

⑷ 死亡3头以上。　　(1539/50000)

17、已知随机变量x服从二项分布B(10,0.6)，求P(2≤x≤6)，P(x≥7)，P(x<3)。

(0.61605，0.38228，0.01229)

18、什么是波松分布？其平均数、方差有何特征？

19、已知随机变量x服从波松分布$P(4)$，求P(x=1)，P(x=2)，P(x≥4)。

(0.0733，0.1465，0.5665)

20、某种疾病的死亡率为0.005。试问在患有此病的360个病例中，(a)有3例或3例以上死亡的概率；(b)恰有3例死亡的概率。　　(0.269，0.160)

21、验收某大批货物时,规定在到货的1000件样品中次品不多于10件时方能接受。如果说整批货物的次品率为0.5％，试求拒收这批货物的概率。(0.014)

22、什么是标准误?标准误与标准差有何联系与区别?

23、样本平均数抽样总体与原始总体的两个参数间有何联系？

24、t分布与标准正态分布有何区别与联系？

第五章　t 检验

前面讲了样本平均数抽样分布的问题。抽样研究的目的是用样本信息来推断总体特征。所谓统计推断是根据样本和假定模型对总体作出的以概率形式表述的推断，它主要包括假设检验（**test of hypothesis**）和参数估计（**parametric estimation**）二个内容。由一个样本平均数可以对总体平均数作出估计，但样本平均数包含有抽样误差，用包含有抽样误差的样本平均数来推断总体，其结论并不是绝对正确的。因而要对样本平均数进行统计假设检验。

假设检验又叫显著性检验（**test of significance**），是统计学中一个很重要的内容。显著性检验的方法很多，常用的有 t 检验、F 检验和 χ^2 检验等。尽管这些检验方法的用途及使用条件不同，但其检验的基本原理是相同的。本章以两个平均数的差异显著性检验为例来阐明显著检验的原理，介绍几种 t 检验的方法，然后介绍总体参数的区间估计（**interval estimation**）。

第一节　显著性检验的基本原理

一、显著性检验的意义

为了便于理解，我们结合一个具体例子来说明显著性检验的意义。随机抽测 10 头长白猪和 10 头大白猪经产母猪的产仔数，资料如下：

长白：11，11，9，12，10，13，13，8，10，13

大白：8，11，12，10，9，8，8，9，10，7

经计算，得长白猪 10 头经产母猪产仔平均数 $\bar{x}_1$=11 头，标准差 S_1=1.76 头；大白猪 10 头经产母猪产仔平均数 $\bar{x}_2$ =9.2 头，标准差 S_2=1.549 头。能否仅凭这两个平均数的差值 $\bar{x}_1-\bar{x}_2$ =1.8 头，立即得出长白与大白两品种经产母猪产仔数不同的结论呢？统计学认为，这样得出的结论是不可靠的。这是因为如果我们再分别随机抽测 10 头长白猪和 10 头大白猪经产母猪的产仔数，又可得到两个样本资料。由于抽样误差的随机性，两样本平均数就不一定是 11 头和 9.2 头，其差值也不一定是 1.8 头。造成这种差异可能有两种原因，一是品种造成的差异，即是长白猪与大白猪本质不同所致，另一可能是试验误差（或抽样误差）。对两个样本进行比较时，必须判断样本间差异是抽样误差造成的，还是本质不同引起的。如何区分两类性质的差异？怎样通过样本来推断总体？这正是显著性检验要解决的问题。

两个总体间的差异如何比较？一种方法是研究整个总体，即由总体中的所有个体数据计算出总体参数进行比较。这种研究整个总体的方法是很准确的，但常常是不可能进行的，因为总体往往是无限总体，或者是包含个体很多的有限总体。因此，不得不采用另一种方法，即研究样本，通过样本研究其所代表的总体。例如，设长白猪经产母猪产仔数的总体平均数为 μ_1，大白猪经产母猪产仔数的总体平均数为 μ_2，试验研究的目的，就是要给 μ_1、μ_2 是否相同做出推断。由于总体平均数 μ_1 、μ_2 未知，在进行显著性检验时只能以样本平均数 $\bar{x}_1$、$\bar{x}_2$ 作为检验对象，更确切地说，是以（$\bar{x}_1-\bar{x}_2$）作为检验对象。

为什么以样本平均数作为检验对象呢？这是因为样本平均数具有下述特征：

1、离均差的平方和Σ（$x-\bar{x}$）^{2}最小。说明样本平均数与样本各个观测值最接近，平均数是资料的代表数。

2、样本平均数是总体平均数的无偏估计值，即E（$\bar{x}$）=μ。

3、根据统计学中心极限定理，样本平均数$\bar{x}$服从或逼近正态分布。

所以，以样本平均数作为检验对象，由两个样本平均数差异的大小去推断样本所属总体平均数是否相同是有其依据的。

由上所述,一方面我们有依据由样本平均数$\bar{x}_1$和$\bar{x}_2$的差异来推断总体平均数μ_1、μ_2相同与否，另一方面又不能仅据样本平均数表面上的差异直接作出结论，其根本原因在于试验误差（或抽样误差）的不可避免性。若对样本观测值的数据结构作一简单剖析，就可更清楚地看到这一点。

通过试验测定得到的每个观测值x_i，既由被测个体所属总体的特征决定，又受个体差异和诸多无法控制的随机因素的影响。所以观测值x_i由两部分组成，即$x_i=\mu+\varepsilon_i$。总体平均数μ反映了总体特征，ε_i表示误差。若样本含量为n，则可得到n个观测值：x_1，x_2，…，x_n。于是样本平均数$\bar{x}=\sum x_i/n=\sum(\mu+\varepsilon_i)/\ n=\mu+\bar{\varepsilon}_i$。说明样本平均数并非总体平均数，它还包含试验误差的成分。

对于接受不同处理的两个样本来说，则有：$\bar{x}_1=\mu_1+\bar{\varepsilon}_1$，$\bar{x}_2=\mu_2+\bar{\varepsilon}_2$。

这说明两个样本平均数之差（$\bar{x}_1-\bar{x}_2$）也包括了两部分：一部分是两个总体平均数的差（$\mu_1-\mu_2$），叫做试验的处理效应**（treatment effect）**；另一部分是试验误差（$\bar{\varepsilon}_1-\bar{\varepsilon}_2$）。也就是说样本平均数的差（$\bar{x}_1-\bar{x}_2$）包含有试验误差，它只是试验的表面效应。因此，仅凭（$\bar{x}_1-\bar{x}_2$）就对总体平均数μ_1、μ_2是否相同下结论是不可靠的。只有通过显著性检验才能从（$\bar{x}_1-\bar{x}_2$）中提取结论。对（$\bar{x}_1-\bar{x}_2$）进行显著性检验就是要分析：试验的表面效应（$\bar{x}_1-\bar{x}_2$）主要由处理效应（$\mu_1-\mu_2$）引起的，还是主要由试验误差所造成。虽然处理效应（$\mu_1-\mu_2$）未知，但试验的表面效应是可以计算的，借助数理统计方法可以对试验误差作出估计。所以，可从试验的表面效应与试验误差的权衡比较中间接地推断处理效应是否存在，这就是显著性检验的基本思想。

为了通过样本对其所在的总体作出符合实际的推断，要求合理进行试验设计，准确地进行试验与观察记载，尽量降低试验误差，避免系统误差，使样本尽可能代表总体。只有从正确、完整而又足够的资料中才能获得可靠的结论。若资料中包含有较大的试验误差与系统误差，有许多遗漏、缺失甚至错误，再好的统计方法也无济于事。因此，收集到正确、完整而又足够的资料是通过显著性检验获得可靠结论的基本前提。

二、显著性检验的基本步骤

仍以前面所举实例说明显著性检验的基本步骤。

（一）首先对试验样本所在的总体作假设 这里假设$\mu_1=\mu_2$或$\mu_1-\mu_2=0$，即假设长白猪和大白猪两品种经产母猪产仔数的总体平均数相等，其意义是试验的表面效应：$\bar{x}_1-\bar{x}_2$=1.8头是试验误差，处理无效，这种假设称为无效假设**（null hypothesis）**，记作H_0：$\mu_1=\mu_2$或$\mu_1-\mu_2=0$。无效假设是被检验的假设，通过检验可能被接受，也可能被否定。提

出 H_0：$\mu_1=\mu_2$ 或 $\mu_1-\mu_2=0$ 的同时，相应地提出一对应假设，称为备择假设（**alternative hypothesis**），记作 H_A。备择假设是在无效假设被否定时准备接受的假设。本例的备择假设是 H_A：$\mu_1\neq\mu_2$ 或 $\mu_1-\mu_2\neq0$，即假设长白猪与大白猪两品种经产母猪产仔数的总体平均数 μ_1 与 μ_2 不相等或 μ_1 与 μ_2 之差不等于零，亦即存在处理效应，其意义是指试验的表面效应，除包含试验误差外，还含有处理效应在内。

（二）在无效假设成立的前提下，构造合适的统计量，并研究试验所得统计量的抽样分布，计算无效假设正确的概率 对于上述例子，研究在无效假设 H_0：$\mu_1=\mu_2$ 成立的前提下，统计量（$\bar{x}_1-\bar{x}_2$）的抽样分布。经统计学研究，得到一个统计量 t：

$$t=\frac{\bar{x}_1-\bar{x}_2}{S_{\bar{x}_1-\bar{x}_2}}$$

其中 $$S_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\sum(x_1-\bar{x}_1)^2+\sum(x_2-\bar{x}_2)^2}{(n_1-1)+(n_2-1)}\times(\frac{1}{n_1}+\frac{1}{n_2})}$$

$S_{\bar{x}_1-\bar{x}_2}$ 叫做均数差异标准误；n_1、n_2 为两样本的含量。

所得的统计量 t 服从自由度 df =（n_1-1）+(n_2-1)的 t 分布。

根据两个样本的数据，计算得：$\bar{x}_1-\bar{x}_2$ =11-9.2=1.8；

$$S_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\sum(x_1-\bar{x}_1)^2+\sum(x_2-\bar{x}_2)^2}{(n_1-1)+(n_2-1)}\times(\frac{1}{n_1}+\frac{1}{n_2})}$$

$$=\sqrt{\frac{28+21.6}{(10-1)+(10-1)}\times(\frac{1}{10}+\frac{1}{10})}=0.742$$

$$t=\frac{\bar{x}_1-\bar{x}_2}{S_{\bar{x}_1-\bar{x}_2}}=\frac{11-9.2}{0.742}=2.426$$

我们需进一步估计出|t|≥2.426 的两尾概率，即估计 P（|t|≥2.426）是多少？查附表 3，在 df=（n_1-1）+(n_2-1)=（10-1）+（10-1）=18 时，两尾概率为 0.05 的临界 t 值：$t_{0.05(18)}$ =2.101，两尾概率为 0.01 的临界 t 值：$t_{0.01(18)}$ =2.878，即：

$$P（|t|>2.101）=P（t>2.101）+P（t<-2.101）=0.05$$

$$P（|t|>2.878）=P（t>2.878）+P（t<-2.878）=0.01$$

由于根据两样本数据计算所得的 t 值为 2.426，介于两个临界 t 值之间，即：

$$t_{0.05}<2.426<t_{0.01}$$

所以，|t|≥2.426 的概率 P 介于 0.01 和 0.05 之间，即：0.01 <P< 0.05。

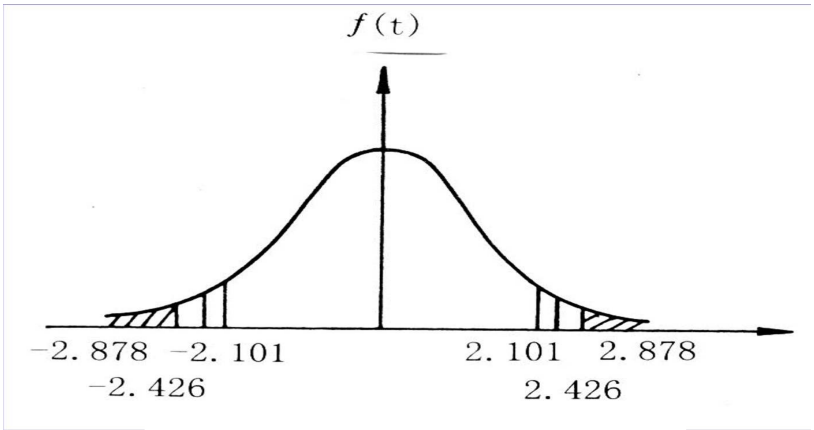


图 5-1 $|t|\geqslant 2.426$ 的两尾概率

如图 5-1 所示，说明无效假设成立的可能性，即试验的表面效应为试验误差的可能性在 0.01—0.05 之间。

（三）根据“小概率事件实际不可能性原理”否定或接受无效假设 上章曾论及：若随机事件的概率很小，例如小于 0.05，0.01，0.001，称之为小概率事件；在统计学上，把小概率事件在一次试验中看成是实际上不可能发生的事件，称为小概率事件实际不可能原理。根据这一原理，当试验的表面效应是试验误差的概率小于 0.05 时，可以认为在一次试验中试验表面效应是试验误差实际上是不可能的，因而否定原先所作的无效假设 H_0：$\mu_1=\mu_2$，接受备择假设 H_A：$\mu_1\neq\mu_2$，即认为：试验的处理效应是存在的。当试验的表面效应是试验误差的概率大于 0.05 时，则说明无效假设 H_0：$\mu_1=\mu_2$ 成立的可能性大，不能被否定，因而也就不能接受备择假设 H_A：$\mu_1\neq\mu_2$。

本例中，按所建立的 H_0：$\mu_1=\mu_2$，试验的表面效应是试验误差的概率在 0.01—0.05 之间，小于 0.05，故有理由否定 H_0：$\mu_1=\mu_2$，从而接受 H_A：$\mu_1\neq\mu_2$。可以认为长白猪与大白猪两品种经产母猪产仔数总体平均数 μ_1 和 μ_2 不相同。

综上所述，显著性检验，从提出无效假设与备择假设到根据小概率事件实际不可能性原理来否定或接受无效假设，这一过程实际上是应用所谓“概率性质的反证法”对试验样本所属总体所作的无效假设的统计推断。对于各种显著性检验的方法，除明确其应用条件，掌握有关统计运算方法外，正确的统计推断是不可忽视的。

三、显著水平与两种类型的错误

在显著性检验中，否定或接受无效假设的依据是“小概率事件实际不可能性原理”。用来确定否定或接受无效假设的概率标准叫显著水平（**significance level**），记作 α。在生物学研究中常取 α=0.05 或 α=0.01。对于上述例子所用的检验方法（t 检验）来说，若

$|t|<t_{0.05}$，则说明试验的表面效应属于试验误差的概率 $P>0.05$，即表面效应属于试验误差的可能性大，不能否定 H_0：$\mu_1=\mu_2$，统计学上把这一检验结果表述为：“两个总体平均数 μ_1 与 μ_2 差异不显著”，在计算所得的 t 值的右上方标记“ns”或不标记符号；若 $t_{0.05}\leqslant|t|<t_{0.01}$，则说明试验的表面效应属于试验误差的概率 P 在 0.01—0.05 之间，即 $0.01<P\leq 0.05$，表面效应属于试验误差的可能性较小，应否定 H_0：$\mu_1=\mu_2$，接受 H_A：$\mu_1\neq\mu_2$，统计学上把这一检验结果表述为：“两个总体平均数 μ_1 与 μ_2 差异显著”，在计算所得的 t 值的右上方标记“*”；若 $|t|\geqslant t_{0.01}$，则说明试验的表面效应属于试验误差的概率 P 不超过 0.01，即 $P\leqslant 0.01$，表面效应属于试验误差的可能性更小，应否定 H_0：$\mu_1=\mu_2$，接受 H_A：$\mu_1\neq\mu_2$，统计学上把这一检验结果表述为：“两个总体平均数 μ_1 与 μ_2 差异极显著”，在计算所得的 t 值的右上方标记“* *”。

这里可以看到，是否否定无效假设 H_0：$\mu_1=\mu_2$，是用实际计算出的检验统计量 t 的绝对值与显著水平 α 对应的临界 t 值 t_α 比较。若 $|t|\geqslant t_\alpha$，则在 α 水平上否定 H_0：$\mu_1=\mu_2$；若 $|t|<t_\alpha$，则不能在 α 水平上否定 H_0：$\mu_1=\mu_2$。区间 $(-\infty,t_\alpha]$ 和 $[t_\alpha,+\infty)$ 称为 α 水平上的否定域，而区间（$-t_\alpha$，t_α）则称为 α 水平上的接受域。

假设检验时选用的显著水平，除 α=0.05 和 0.01 为常用外，也可选 α=0.10 或 α=0.001 等等。到底选哪种显著水平，应根据试验的要求或试验结论的重要性而定。如果试验中难以控制的因素较多，试验误差可能较大，则显著水平可选低些，即 α 值取大些。反之，如试验耗费较大，对精确度的要求较高，不容许反复，或者试验结论的应用事关重大，则所选显著水平应高些，即 α 值应该小些。显著水平 α 对假设检验的结论是有直接影响的，所以它应在试验开始前即确定下来。

因为显著性检验是根据“小概率事件实际不可能性原理”来否定或接受无效假设的，所以不论是接受还是否定无效假设，都没有 100%的把握。也就是说，在检验无效假设 H_0 时可能犯两类错误。第一类错误是真实情况为 H_0 成立，却否定了它，犯了“弃真”错误，也叫Ⅰ型错误（**type Ⅰ error**）。Ⅰ型错误，就是把非真实差异错判为真实差异，即 H_0：$\mu_1=\mu_2$ 为真，却接受了 H_A：$\mu_1\neq\mu_2$。第二类错误是 H_0 不成立，却接受了它，犯了“纳伪”错误，也叫Ⅱ型错误（**type Ⅱ error**）。Ⅱ型错误，就是把真实差异错判为非真实差异，即 H_A：$\mu_1\neq\mu_2$ 为真，却未能否定 H_0：$\mu_1=\mu_2$。

我们是基于“小概率事件实际不可能性原理”来否定 H_0，但在一次试验中小概率事件并不是绝对不会发生的。如果我们抽得一个样本，它虽然来自与 H_0 对应的抽样总体，但计算所得的统计量 t 却落入了否定域中，因而否定了 H_0，于是犯了Ⅰ型错误。但犯这类错误的概率不会超过 α。

Ⅱ型错误发生的原因可以用图 5-2 来说明。图中左边曲线是 H_0：$\mu_1=\mu_2$ 为真时，（$\bar{x}_1-\bar{x}_2$）的分布密度曲线；右边曲线是 H_A：$\mu_1\neq\mu_2$ 为真时，（$\bar{x}_1-\bar{x}_2$）的分布密度曲线；右边曲线是 H_A：$\mu_1\neq\mu_2$ 为真时，（$\bar{x}_1-\bar{x}_2$）的分布密度曲线（$\mu_1>\mu_2$），它们构成的抽样

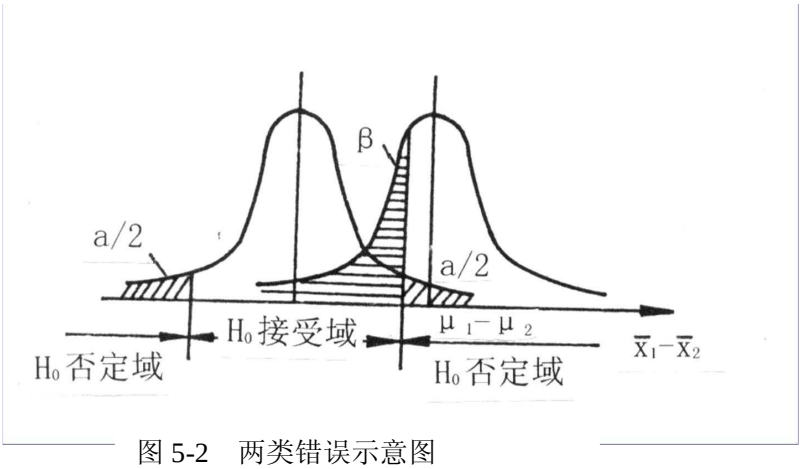


图 5-2　两类错误示意图

分布相叠加。有时我们从 $\mu_1-\mu_2\neq0$ 抽样总体抽取一个（ $\bar{x}_1-\bar{x}_2$ ）恰恰在 H_0 成立时的接受域内（如图中横线阴影部分），这样，实际是从 $\mu_1-\mu_2\neq0$ 总体抽的样本，经显著性检验却不能否定 H_0，因而犯了Ⅱ型错误。犯Ⅱ型错误的概率用 β 表示。 误概率 β 值的大小较难确切估计，它只有与特定的 H_A 结合起来才有意义。一般与显著水平 α、原总体的标准差σ、样本含量 n、以及相互比较的两样本所属总体平均数之差 $\mu_1-\mu_2$ 等因素有关。在其它因素确定时， α 值越小， β 值越大；反之， α 值越大， β 值越小；样本含量 n 及 $\mu_1-\mu_2$ 越大、σ越小， β 值越小。

由于 β 值的大小与 α 值的大小有关，所以在选用检验的显著水平时应考虑到犯Ⅰ、Ⅱ型错误所产生后果严重性的大小，还应考虑到试验的难易及试验结果的重要程度。若一个试验耗费大，可靠性要求高，不允许反复，那么 α 值应取小些；当一个试验结论的使用事关重大，容易产生严重后果，如药物的毒性试验， α 值亦应取小些。对于一些试验条件不易控制，试验误差较大的试验，可将 α 值放宽到 0.1，甚至放宽到 0.25。

在提高显著水平，即减小 α 值时，为了减小犯Ⅱ型错误的概率，可适当增大样本含量。因为增大样本含量可使（ $\bar{x}_1-\bar{x}_2$ ）分布的方差 σ^2（ $1/n_1+1/n_2$ ）变小，使图 5-2 左右两曲线变得比较“高”、“瘦”，叠加部分减少，即 β 值变小。我们的愿望是 α 值不越过某个给定值，比如 α=0.05 或 0.01 的前提下， β 值越小越好。因为在具体问题中 $\mu_1-\mu_2$ 和σ相对不变，所以 β 值的大小主要取决于样本含量的大小。

图 5-2 中的 $1-\beta$ 称为检验功效或检验力（**power of test**），也叫把握度。其意义是当两总体确有差别（即 H_A 成立）时，按 α 水平能发现它们有差别的能力。例如 $1-\beta$=0.9，意味着若两总体确有差别，则理论上平均 100 次抽样比较中有 90 次能得出有差别的结论。

两类错误的关系可归纳如下：

表 5-1　两类错误的关系

客观实际	否定 H_0	接受 H_0
H_0 成立	Ⅰ型错误（α）	推断正确（1-α）
H_0 不成立	推断正确（1-β）	Ⅱ型错误（β）

四、双侧检验与单侧检验

在上述显著性检验中，无效假设 H_0：$\mu_1=\mu_2$ 与备择假设 H_A：$\mu_1\neq\mu_2$。此时，备择假设中包括了 $\mu_1>\mu_2$ 或 $\mu_1<\mu_2$ 两种可能。这个假设的目的在于判断 μ_1 与 μ_2 有无差异，而不考虑谁大谁小。如比较长白猪与大白猪两品种猪经产母猪的产仔数，长白猪可能高于大白猪，也可能低于大白猪。

此时，在 α 水平上否定域为 $(-\infty,t_\alpha]$ 和 $[t_\alpha,+\infty)$，对称地分配在 t 分布曲线的两侧尾部，每侧的概率为 $\alpha/2$，如图 5-3 所示。这种利用两尾概率进行的检验叫双侧检验（**two-sided test**），也叫双尾检验（**two-tailed test**），t_α 为双侧检验的临界 t 值。但在有些情况下，双侧检验不一定符合实际情况。如采用某种新的配套技术措施以期提高鸡的产蛋量，已知此种配套技术的实施不会降低产蛋量。此时，若进行新技术与常规技术的比较试验，则无效假设应为 H_0：$\mu_1=\mu_2$，即假设新技术与常规技术产蛋量是相同的，备择假设应为 H_A：$\mu_1>\mu_2$，即新配套技术的实施使产蛋量有所提高。检验的目的在于推断实施新技术是否提高了产蛋量，这时 H_0 的否定域在 t 分布曲线的右尾。在 α 水平上否定域为 $[t_\alpha,+\infty)$，右侧的概率为 α，如图 5-4*A* 所示。若无效假设 H_0 为 $\mu_1=\mu_2$，备择假设 H_A 为 $\mu_1<\mu_2$，此时 H_0 的否定域在 t 分布曲线的左尾。在 α 水平上，H_0 的否定域为 $(-\infty,-t_\alpha]$，左侧的概率为 α。如图 5-4*B* 所示。这种利用一尾概率进行的检验叫单侧检验（**one-sided test**）也叫单尾检验（**one-tailed test**）。此时 t_α 为单侧检验的临界 t 值。显然，单侧检验的 t_α=双侧检验的 $t_{2\alpha}$。

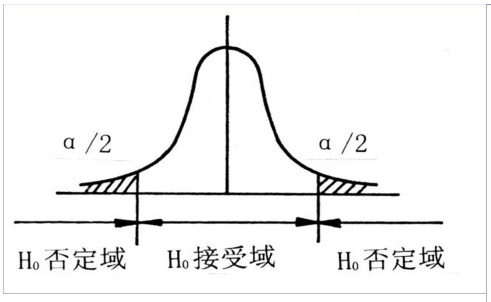


图 5-3　双侧检验

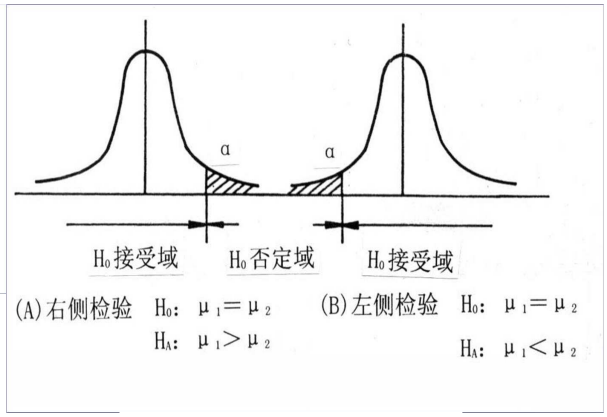


图 5-4　单侧检验

由上可以看出，若对同一资料进行双侧检验也进行单侧检验，那么在 α 水平上单侧检验显著，只相当于双侧检验在 2α 水平上显著。所以，同一资料双侧检验与单侧检验所得的结论不一定相同。双侧检验显著，单侧检验一定显著；但单侧检验显著，双侧检验未必

显著。

选用单侧检验还是双侧检验应根据专业知识及问题的要求在试验设计时就确定。一般若事先不知道所比较的两个处理效果谁好谁坏，分析的目的在于推断两个处理效果有无差别，则选用双侧检验；若根据理论知识或实践经验判断甲处理的效果不会比乙处理的效果差（或相反），分析的目的在于推断甲处理是否比乙处理好（或差），则用单侧检验。一般情况下，如不作特殊说明均指双侧检验。

五、显著性检验中应注意的问题

上面我们已详细阐明了显著性检验的意义及原理。进行显著性检验还应注意以下几个问题：

（一）为了保证试验结果的可靠及正确，要有严密合理的试验或抽样设计，保证各样本是从相应同质总体中随机抽取的。并且处理间要有可比性，即除比较的处理外，其它影响因素应尽可能控制相同或基本相近。否则，任何显著性检验的方法都不能保证结果的正确。

（二）选用的显著性检验方法应符合其应用条件。上面我们所举的例子属于“非配对设计两样本平均数差异显著性检验”。由于研究变量的类型、问题的性质、条件、试验设计方法、样本大小等的不同，所用的显著性检验方法也不同，因而在选用检验方法时，应认真考虑其适用条件，不能滥用。

（三）要正确理解差异显著或极显著的统计意义。显著性检验结论中的“差异显著”或“差异极显著”不应该误解为相差很大或非常大，也不能认为在专业上一定就有重要或很重要的价值。“显著”或“极显著”是指表面上如此差别的不同样本来自同一总体的可能性小于 0.05 或 0.01，已达到了可以认为它们有实质性差异的显著水平。有些试验结果虽然差别大，但由于试验误差大，也许还不能得出“差异显著”的结论，而有些试验的结果间的差异虽小，但由于试验误差小，反而可能推断为“差异显著”。

显著水平的高低只表示下结论的可靠程度的高低，即在 0.01 水平下否定无效假设的可靠程度为 99％，而在 0.05 水平下否定无效假设的可靠程度为 95%。

“差异不显著”是指表面上的这种差异在同一总体中出现的可能性大于统计上公认的概率水平 0.05，不能理解为试验结果间没有差异。下“差异不显著”的结论时，客观上存在两种可能：一是本质上有差异，但被试验误差所掩盖，表现不出差异的显著性来。如果减小试验误差或增大样本含量，则可能表现出差异显著性；二是可能确无本质上差异。显著性检验只是用来确定无效假设能否被推翻，而不能证明无效假设是正确的。

（四）合理建立统计假设，正确计算检验统计量。就两个样本平均数差异显著性检验来说，无效假设 H_0 与备择假设 H_A 的建立，一般如前所述，但也有时也例外。如经收益与成本的综合经济分析知道，饲喂畜禽以高质量的Ⅰ号饲料比饲喂Ⅱ号饲料提高的成本需用畜禽生产性能提高 d 个单位获得的收益来相抵，那么在检验喂Ⅰ号饲料与Ⅱ号饲料在收益上是否有差异时，无效假设应为 H_0：$\mu_1-\mu_2=d$，备择假设为 H_A：$\mu_1-\mu_2\neq d$（双侧检验）；或 H_A：$\mu_1-\mu_2>d$（单侧检验）；t 检验计算公式为：

$$t=\frac{(\bar{x}_1-\bar{x}_2)-d}{S_{\bar{x}_1-\bar{x}_2}} \qquad (5\text{-}1)$$

如果不能否定无效假设，可以认为喂高质量的Ⅰ号饲料得失相抵，只有当（$\bar{x}_1-\bar{x}_2$）$>d$ 达到一定程度而否定了 H_0，才能认为喂Ⅰ号饲料可获得更多的收益。

(五)结论不能绝对化。经过显著性检验最终是否否定无效假设则由被研究事物有无本质差异、试验误差的大小及选用显著水平的高低决定的。同样一种试验，试验本身差异程度的不同，样本含量大小的不同，显著水平高低的不同，统计推断的结论可能不同。否定 H_0 时可能犯Ⅰ型错误，接受 H_0 时可能犯Ⅱ型错误。尤其在 P 接近 α 时，下结论应慎重，有时应用重复试验来证明。总之，具有实用意义的结论要从多方面综合考虑，不能单纯依靠统计结论。

此外，报告结论时应列出，由样本算得的检验统计量值（如 t 值），注明是单侧检验还是双侧检验，并写出 P 值的确切范围，如 $0.01<P<0.05$，以便读者结合有关资料进行对比分析。

第二节　样本平均数与总体平均数差异显著性检验

在实际工作中我们往往需要检验一个样本平均数与已知的总体平均数是否有显著差异，即检验该样本是否来自某一总体。已知的总体平均数一般为一些公认的理论数值、经验数值或期望数值。如畜禽正常生理指标、怀孕期、家禽出雏日龄以及生产性能指标等，都可以用样本平均数与之比较，检验差异显著性。检验的基本步骤是：

（一）提出无效假设与备择假设　H_0：$\mu=\mu_0$，H_A：$\mu\neq\mu_0$，其中 μ 为样本所在总体平均数，μ_0 为已知总体平均数；

（二）计算 t 值　计算公式为：

$$t=\frac{\bar{x}-\mu_0}{S_{\bar{x}}} \qquad df=n-1 \qquad (5\text{-}2)$$

式中，n 为样本含量，$S_{\bar{x}}=\frac{S}{\sqrt{n}}$ 为样本标准误。

（三）查临界 t 值，作出统计推断　由 $df=n-1$ 查附表 3 得临界值 $t_{0.05}$，$t_{0.01}$。将计算所得 t 值的绝对值与其比较，若 $|t|<t_{0.05}$，则 $P>0.05$，不能否定 H_0：$\mu=\mu_0$，表明样本平均数 $\bar{x}$ 与总体平均数 μ_0 差异不显著，可以认为样本是取自该总体；若 $t_{0.05}\leqslant|t|<t_{0.01}$，则 $0.01<P\leqslant0.05$，否定 H_0：$\mu=\mu_0$，接受 H_A：$\mu\neq\mu_0$，表明样本平均数 $\bar{x}$ 与总体平均数 μ_0 差异显著，有 95%的把握认为样本不是取自该总体；若 $|t|\geqslant t_{0.01}$，则 $P\leqslant0.01$，表明样本平均数 $\bar{x}$ 与总体平均数 μ_0 差异极显著，有 99%的把握认为样本不是取自该总体。

若在 0.05 水平上进行单侧检验，只要将计算所得 t 值的绝对值 $|t|$ 与由附表 3 查得 $\alpha=0.10$ 的临界 t 值 $t_{0.10}$ 比较，即可作出统计推断。

【例 5.1】 母猪的怀孕期为 114 天，今抽测 10 头母猪的怀孕期分别为 116、115、113、112、114、117、115、116、114、113（天），试检验所得样本的平均数与总体平均数 114

天有无显著差异？

根据题意，本例应进行双侧 t 检验。

1、提出无效假设与备择假设　H_0：μ=114 ，H_A：$\mu \neq 114$

2、计算t值

经计算得：$\bar{x}$=114.5，S=1.581

所以　$t=\dfrac{\bar{x}-u_0}{S_{\bar{x}}}=\dfrac{114.5-114}{1.581/\sqrt{10}}=\dfrac{0.5}{0.5}=1.000$

$df=n-1$=10-1=9

3、查临界t值，作出统计推断　由df =9，查t值表（附表3）得$t_{0.05(9)}$ =2.262，因为$|t|<t_{0.05}$，$P>0.05$，故不能否定H_0：μ=114，表明样本平均数与总体平均数差异不显著，可以认为该样本取自母猪怀孕期为114天的总体。

【例5.2】 按饲料配方规定，每1000kg某种饲料中维生素C不得少于246g，现从工厂的产品中随机抽测12个样品，测得维生素C含量如下：255、260、262、248、244、245、250、238、246、248、258、270g/1000kg，若样品的维生素C含量服从正态分布，问此产品是否符合规定要求？

按题意，此例应采用单侧检验。

1、提出无效假设与备择假设　H_0：μ=246，H_A：$\mu>250$

2、计算t值

经计算得：$\bar{x}$=252，S=9.115

所以$t=\dfrac{\bar{x}-u}{S_{\bar{x}}}=\dfrac{252-246}{9.115/\sqrt{12}}=\dfrac{6}{2.631}$=2.281

$df=n-1$=12-1=11

3、查临界t值，作出统计推断　因为单侧$t_{0.05(11)}$ =双侧$t_{0.10(11)}$=1.796，$|t|>$单侧$t_{0.05(11)}$，$P<0.05$，否定H_0：μ=246，接受H_A：$\mu>246$，表明样本平均数与总体平均数差异显著，可以认为该批饲料维生素C含量符合规定要求。

第三节　两个样本平均数的差异显著性检验

在实际工作中还经常会遇到推断两个样本平均数差异是否显著的问题，以了解两样本所属总体的平均数是否相同。对于两样本平均数差异显著性检验，因试验设计不同，一般可分为两种情况：一是非配对设计或成组设计两样本平均数的差异显著性检；二是配对设计两样本平均数的差异显著性检。

一、非配对设计两样本平均数的差异显著性检验

非配对设计或成组设计是指当进行只有两个处理的试验时，将试验单位完全随机地分成两个组，然后对两组随机施加一个处理。在这种设计中两组的试验单位相互独立，所得的二个样本相互独立，其含量不一定相等。非配对设计资料的一般形式见表 5-2。

表 5-2　非配对设计资料的一般形式

处理	观测值 x_{ij}	样本含量 n_i	平均数 $\bar{x}$	总体平均数
1	$x_{11}\quad x_{12}\quad \dots\quad x_{1n_1}$	n_1	$\bar{x}_1=\Sigma x_{1j}/\ n_1$	μ_1
2	$x_{21}\quad x_{22}\quad \dots x_{2n_2}$	n_2	$\bar{x}_2=\Sigma x_{2j}/\ n_2$	μ_2

非配对设计两样本平均数差异显著性检验的基本步骤如下：

（一）提出无效假设与备择假设　H_0：$\mu_1=\mu_2$，H_A：$\mu_1\neq\mu_2$

（二）计算 t 值　计算公式为：

$$t=\frac{\bar{x}_1-\bar{x}_2}{S_{\bar{x}_1-\bar{x}_2}}\qquad df=(n_1-1)+(n_2-1)\tag{5-3}$$

其中：
$$S_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\sum(x_1-\bar{x}_1)^2+\sum(x_2-\bar{x}_2)^2}{(n_1-1)+(n_2-1)}\times\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\tag{5-4}$$

$$=\sqrt{\frac{\left[\sum x_1^{\,2}-\frac{\left(\sum x_1\right)^2}{n_1}\right]+\left[\sum x_2^{\,2}-\frac{\left(\sum x_2\right)^2}{n_2}\right]}{(n_1-1)+(n_2-1)}\times\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$=\sqrt{\frac{(n_1-1)S_1^{\,2}+(n_2-1)S_2^{\,2}}{(n_1-1)+(n_2-1)}\times\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

当 $n_1=n_2=n$ 时，

$$S_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\sum(x_1-\bar{x}_1)^2+\sum(x_2-\bar{x}_2)^2}{n(n-1)}}=\sqrt{\frac{S_1^{\,2}}{n}+\frac{S_2^{\,2}}{n}}=\sqrt{S_{\bar{x}_1}^{\,2}+S_{\bar{x}_2}^{\,2}}\tag{5-5}$$

$S_{\bar{x}_1-\bar{x}_2}$ 为均数差异标准误，n_1、n_2，$\bar{x}_1$、$\bar{x}_2$，$S_1^{\,2}$、$S_2^{\,2}$ 分别为两样本含量、平均数、均方。

（三）根据 $df=(n_1-1)+(n_2-1)$，查临界 t 值：$t_{0.05}$、$t_{0.01}$，将计算所得 t 值的绝对值与其比较，作出统计推断

【例 5.3】 某种猪场分别测定长白后备种猪和蓝塘后备种猪 90kg 时的背膘厚度，测定结果如表 5-3 所示。设两品种后备种猪 90kg 时的背膘厚度值服从正态分布，且方差相等，问该两品种后备种猪 90kg 时的背膘厚度有无显著差异？

表 5-3　长白与蓝塘后备种猪背膘厚度

品种	头数	背膘厚度（cm）
长白	12	1.20、1.32、1.10、1.28、1.35、1.08、1.18、1.25、1.30、1.12、1.19、1.05
蓝塘	11	2.00、1.85、1.60、1.78、1.96、1.88、1.82、1.70、1.68、1.92、1.80

1、提出无效假设与备择假设　H_0：$\mu_1=\mu_2$，H_A：$\mu_1\neq\mu_2$

2、计算t值　此例n_1=12、n_2=11，经计算得$\bar{x}_1$=1.202、S_1=0.0998、SS_1=0.1096，$\bar{x}_2$=1.817、S_2=0.123、SS_2=0.1508

SS_1、SS_2分别为两样本离均差平方和。

$$S_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\sum(x_1-\bar{x}_1)^2+\sum(x_2-\bar{x}_2)^2}{(n_1-1)+(n_2-1)}\times(\frac{1}{n_1}+\frac{1}{n_2})}$$

$$=\sqrt{\frac{0.1096+0.1508}{(12-1)+(11-1)}\times(\frac{1}{12}+\frac{1}{11})}$$

$$=\sqrt{0.00216}$$

$$=0.0465$$

$$t=\frac{\bar{x}_1-\bar{x}_2}{S_{\bar{x}_1-\bar{x}_2}}=\frac{1.202-1.817}{0.0465}=-13.226^{**}$$

$df=(n_1-1)+(n_2-1)$=（12-1）+（11-1）=21

3、查临界t值，作出统计推断　当df=21时，查临界t值得：$t_{0.01(21)}$=2.831，$|t|$>2.831，P<0.01，否定H_0：$\mu_1=\mu_2$，接受H_A：$\mu_1\neq\mu_2$，表明长白后备种猪与蓝塘后备种猪90kg背膘厚度差异极显著，这里表现为长白后备种猪的背膘厚度极显著地低于蓝塘后备种猪的背膘厚度。

【例5.4】 某家禽研究所对粤黄鸡进行饲养对比试验，试验时间为60天，增重结果如表5-4，问两种饲料对粤黄鸡的增重效果有无显著差异？

表5-4　粤黄鸡饲养试验增重

饲料	n_i	增　重（g）
A	8	720、710、735、680、690、705、700、705
B	8	680、695、700、715、708、685、698、688

此例$n_1=n_2=8$，经计算得$\bar{x}_1$=705.625、S_1^2=288.839，$\bar{x}_2$=696.125、S_2^2=138.125

1、提出无效假设与备择假设　H_0：$\mu_1=\mu_2$，H_A：$\mu_1\neq\mu_2$

2、计算t值，

因为　$S_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{S_1^2+S_2^2}{n}}=\sqrt{\frac{288.839+138.125}{8}}$=7.306

于是　$t=\frac{\bar{x}_1-\bar{x}_2}{S_{\bar{x}_1-\bar{x}_2}}=\frac{705.625-696.125}{7.306}$=1.300

$df=(n_1-1)+(n_2-1)$=（8-1）+（8-1）=14

3、查临界t值，作出统计推断　当df=14时，查临界t值得：$t_{0.05(14)}$=2.145，$|t|$<2.145, P>0.05，故不能否定无效假设H_0：$\mu_1=\mu_2$，表明两种饲料饲喂粤黄鸡的增重效果差异不显著，可以认为两种饲料的质量是相同的。

在非配对设计两样本平均数的差异显著性检验中，若总的试验单位数（n_1+n_2）不变，则两样本含量相等比两样本含量不等有较高检验效率，因为此时使$S_{\bar{x}_1-\bar{x}_2}$最小，从而使t的

绝对值最大。所以在进行非配对设计时，两样本含量以相同为好。

二、配对设计两样本平均数的差异显著性检验

非配对设计要求试验单位尽可能一致。如果试验单位变异较大，如试验动物的年龄、体重相差较大，若采用上述方法就有可能使处理效应受到系统误差的影响而降低试验的准确性与精确性。为了消除试验单位不一致对试验结果的影响，正确地估计处理效应，减少系统误差，降低试验误差，提高试验的准确性与精确性，可以利用局部控制的原则，采用配对设计。

配对设计是指先根据配对的要求将试验单位两两配对，然后将配成对子的两个试验单位随机地分配到两个处理组中。配对的要求是，配成对子的两个试验单位的初始条件尽量一致，不同对子间试验单位的初始条件允许有差异，每一个对子就是试验处理的一个重复。配对的方式有两种：自身配对与同源配对。

1、**自身配对**　指同一试验单位在二个不同时间上分别接受前后两次处理，用其前后两次的观测值进行自身对照比较；或同一试验单位的不同部位的观测值或不同方法的观测值进行自身对照比较。如观测某种病畜治疗前后临床检查结果的变化；观测用两种不同方法对畜产品中毒物或药物残留量的测定结果变化等。

2、**同源配对**　指将来源相同、性质相同的两个个体配成一对，如将畜别、品种、窝别、性别、年龄、体重相同的两个试验动物配成一对，然后对配对的两个个体随机地实施不同处理。

在配对设计中，由于各对试验单位间存在系统误差，对内两个试验单位存在相似性，其资料的显著性检验不同于非配对设计。配对设计试验资料的一般形式见表 5-5。

表 5-5　配对设计试验资料的一般形式

处理	观测值 x_{ij}				样本含量	样本平均数	总体平均数
1	x_{11}	x_{12}	·	x_{1n}	n	$\bar{x}_1=\Sigma x_{1j}/n$	μ_1
2	x_{21}	x_{22}	·	x_{2n}	n	$\bar{x}_2=\Sigma x_{2j}/n$	μ_2
$d_j=x_{1j}-x_{2j}$	d_1	d_2	·	d_n	n	$\bar{d}=\bar{x}_1-\bar{x}_2$	$\mu_d=\mu_1-\mu_2$

配对设计两样本平均数差异显著性检验的基本步骤如下：

（一）提出无效假设与备择假设　H_0：$\mu_d=0$，H_A：$\mu_d\neq 0$，其中 μ_d 为两样本配对数据差值 d 总体平均数，它等于两样本所属总体平均数 μ_1 与 μ_2 之差，即 $\mu_d=\mu_1-\mu_2$。所设无效假设、备择假设相当于 H_0：$\mu_1=\mu_2$，H_A：$\mu_1\neq\mu_2$。

（二）计算 t 值　计算公式为:

$$t=\frac{\bar{d}}{S_{\bar{d}}},\quad df=n-1 \tag{5-6}$$

式中，$S_{\bar{d}}$ 为差异标准误，计算公式为:

$$S_{\bar{d}}=\frac{S_d}{\sqrt{n}}=\sqrt{\frac{\sum(d-\bar{d})^2}{n(n-1)}}=\sqrt{\frac{\sum d^2-(\sum d)^2/n}{n(n-1)}} \tag{5-7}$$

d 为两样本各对数据之差：$d_j = x_{1j} - x_{2j}$，$(j = 1,2,\ldots,n)$；$\bar{d} = \sum d_j / n$；S_d 为 d 的标准差；n 为配对的对子数，即试验的重复数。

（三）查临界 t 值，作出统计推断　根据 $df = n-1$ 查临界 t 值：$t_{0.05(n-1)}$ 和 $t_{0.01(n-1)}$，将计算所得 t 值的绝对值与其比较，作出推断。

【例 5.5】 用家兔 10 只试验某批注射液对体温的影响，测定每只家兔注射前后的体温，见表 5-6。设体温服从正态分布，问注射前后体温有无显著差异？

表 5-6　10 只家兔注射前后的体温

兔号	1	2	3	4	5	6	7	8	9	10
注射前体温	37.8	38.2	38.0	37.6	37.9	38.1	38.2	37.5	38.5	37.9
注射后体温	37.9	39.0	38.9	38.4	37.9	39.0	39.5	38.6	38.8	39.0
$d = x_1 - x_2$	-0.1	-0.8	-0.9	-0.8	0	-0.9	-1.3	-1.1	-0.3	-1.1

1、提出无效假设与备择假设

H_0：μ_d=0，即假定注射前后体温无差异

H_A：$\mu_d \neq 0$，即假定注射前后体温有差异

2、计算 t 值　经过计算得 $\bar{d}$ =-0.73，$S_{\bar{d}} = S_d / \sqrt{n} = 0.445/\sqrt{10} = 0.141$

故　$$t = \frac{\bar{d}}{S_{\bar{d}}} = \frac{-0.73}{0.141} = -5.177$$

且　$df = n-1$=10-1=9

3、查临界 t 值，作出统计推断　由 df =9,查 t 值表得：$t_{0.01(9)}$ =3.250，$|t| > t_{0.01(9)}$，P<0.01，否定 H_0：μ_d=0，接受 H_A：$\mu_d \neq 0$，表明家兔注射该批注射液前后体温差异极显著，注射该批注射液可使体温极显著升高。

【例 5.6】 现从 8 窝仔猪中每窝选出性别相同、体重接近的仔猪两头进行饲料对比试验，将每窝两头仔猪随机分配到两个饲料组中，时间 30 天，试验结果见表 5-7。问两种饲料喂饲仔猪增重有无显著差异？

表 5-7　仔猪饲料对比试验　　单位：*kg*

窝号	1	2	3	4	5	6	7	8
甲饲料（x_1）	10.0	11.2	11.0	12.1	10.5	9.8	11.5	10.8
乙饲料（x_2）	9.8	10.6	9.0	10.5	9.6	9.0	10.8	9.8
$d = x_1 - x_2$	0.2	0.6	2.0	1.6	0.9	0.8	0.7	1.0

1、提出无效假设与备择假设

H_0：μ_d=0，即假定两种饲料喂饲仔猪平均增重无差异

H_A：$\mu_d \neq 0$，即假定两种饲料喂饲仔猪平均增重有差异

2、计算 t 值　计算得　$\bar{d}$ =0.975，$S_{\bar{d}} = S_d / \sqrt{n} = 0.5726/\sqrt{8} = 0.2025$

故　$$t = \frac{\bar{d}}{S_{\bar{d}}} = \frac{-0.975}{0.2025} = 4.815$$

且　$df = n-1$=8-1=7

3、查临界 t 值，作出统计推断　由 df=7，查 t 值表得：$t_{0.01(7)}$=3.499，$|t|$>3.499，P<0.01，表明甲种饲料与乙种饲料喂饲仔猪平均增重差异极显著，这里表现为甲种饲料喂饲仔猪的平均增重极显著高于乙种饲料喂饲的仔猪平均增重。

一般说来，相对于非配对设计，配对设计能够提高试验的精确性。两种方法估计误差的公式为：

非配对设计（$n_1 = n_2 = n$）

$$S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sum(x_1-\bar{x}_1)^2+\sum(x_2-\bar{x}_2)^2}{n(n-1)}}$$

配对设计

$$S_{\bar{d}} = \sqrt{\frac{\sum(d-\bar{d})^2}{n(n-1)}}$$

两式中被开方表达式的分母相同。

因为 $\sum(d_j-\bar{d})^2 = \sum[(x_{1j}-x_{2j})-(\bar{x}_1-\bar{x}_2)]^2 = \sum[(x_{1j}-\bar{x}_1)-(x_{2j}-\bar{x}_2)]^2$

$$= \sum(x_{1j}-\bar{x}_1)^2+\sum(x_{2j}-\bar{x}_2)^2-2\sum(x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2)$$

在配对设计中，$(x_{1j}-\bar{x}_1)$ 和 $(x_{2j}-\bar{x}_2)$ 有同时为正和同时为负的倾向，故 $\sum(x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2)$ 常大于 0，$\sum(d_j-\bar{d})^2$ 常小于 $\sum[(x_{1j}-\bar{x}_1)^2+(x_{2j}-\bar{x}_2)^2]$，这样 $S_{\bar{d}}$ 常小于 $S_{\bar{x}_1-\bar{x}_2}$。但并非所有 $\sum(x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2)$ 恒大于零，故有时 $S_{\bar{d}}$ 大于 $S_{\bar{x}_1-\bar{x}_2}$，也就是说并非所有的配对设计的试验误差都小于非配对设计的试验误差。这就要求我们在进行配对设计时，配成对子的两个试验单位必须真正符合配对条件，若试验单位不具备配对条件，不要勉强采用配对设计。

此外，还须指出，因为配对设计误差自由度为非配对设计（n_1=n_2=n）误差自由度的一半，使得配对设计的临界 t 值大于非配对设计的临界 t 值，于是配对设计只有因 $S_{\bar{d}}$ 的减小而使计算的 t 的绝对值增大的程度超过因自由度减小而使临界 t 值增大的程度，才能比非配对设计更有效地发现两样本间的真实差异。

在进行两样本平均数差异显著性检验时，亦有双侧与单侧检验之分。关于单侧检验，只要注意问题的性质、备择假设 H_A 的建立和临界 t 值的查取就行了，具体计算与双侧检验相同。

第三节　百分数资料差异显著性检验

在第四章介绍二项分布时曾指出：由具有两个属性类别的质量性状利用统计次数法得来的次数资料进而计算出的百分数资料，如成活率、死亡率、孵化率、感染率、阳性率等

是服从二项分布的。这类百分数的假设检验应按二项分布进行。当样本含量 n 较大，p 不过小，且 np 和 nq 均大于 5 时，二项分布接近于正态分布。所以，对于服从二项分布的百分数资料，当 n 足够大时，可以近似地用 u 检验法，即自由度为无穷大时（$df=\infty$）的 t 检验法，进行差异显著性检验。适用于近似地采用 u 检验所需的二项分布百分数资料的样本含量 n 见表 5-8。

表 5-8　适用于近似地采用 u 检验所需要的二项分布百分数资料的样本含量 n

$\hat{p}$（样本百分数）	$n\hat{p}$（较小百分数的次数）	n（样本含量）
0.5	15	30
0.4	20	50
0.3	24	80
0.2	40	200
0.1	60	600
0.05	70	1,400

与平均数差异显著性检验类似，百分数差异显著性检验分为样本百分数与总体百分数差异显著性检验及两样本百分数差异显著性检验两种。

一、样本百分数与总体百分数差异显著性检验

在实际工作中，有时需要检验一个服从二项分布的样本百分数与已知的二项总体百分数差异是否显著，其目的在于检验一个样本百分数 $\hat{p}$ 所在二项总体百分数 p 是否与已知二项总体百分数 p_0 相同，换句话说，检验该样本百分数 $\hat{p}$ 是否来自总体百分数为 p_0 的二项总体。这里所讨论的百分数是服从二项分布的，但 n 足够大，p 不过小，np 和 nq 均大于 5，可近似地采用 u 检验法来进行显著性检验；若 np 或 nq 小于或等于 30 时，应对 u 进行连续性矫正。检验的基本步骤是：

（一）提出无效假设与备择假设

$H_0 : p = p_0$，　$H_A : p \neq p_0$

（二）计算 u 值或 u_c 值

u 值的计算公式为：

$$u = \frac{\hat{p} - p_0}{S_{\hat{p}}} \tag{5-8}$$

矫正 u 值 u_c 的计算公式为：

$$u_c = \frac{|\hat{p} - p_0| - 0.5/n}{S_{\hat{p}}} \tag{5-9}$$

其中 $\hat{p}$ 为样本百分数，p_0 为总体百分数，$S_{\hat{p}}$ 为样本百分数标准误，计算公式为：

$$S_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} \tag{5-10}$$

（三）将计算所得的 u 或 u_c 的绝对值与 1.96、2.58 比较，作出统计推断 若 $|u|$（或 $|u_c|$）<1.96,p>0.05，不能否定 $H_0: p = p_0$，表明样本百分数 $\hat{p}$ 与总体百分数 p_0 差异不显著；若 $1.96 \le |u|$(或$|u_c|$) <2.58，0.01<p≤0.05，否定 $H_0: p = p_0$，接受 $H_A: p \ne p_0$，表明样本百分数 $\hat{p}$ 与总体百分数 p_0 差异显著；若 $|u|$(或$|u_c|$) ≥ 2.58，$p \le 0.01$，否定 $H_0: p = p_0$，接受 $H_A: p \ne p_0$，表明样本百分数 $\hat{p}$ 与总体百分数 p_0 差异极显著。

【例 5.7】 据往年调查某地区的乳牛隐性乳房炎一般为 30%，现对某牛场 500 头乳牛进行检测，结果有 175 头乳牛凝集反应阳性，问该牛场的隐性乳房炎是否比往年严重？

此例总体百分数 p_0=30%，样本百分数 $\hat{p}$ =175/500=35%，因为 $np_0 = 500 \times 30\%$ =150>30，不须进行连续性矫正。

1、提出无效假设与备择假设 $H_0: p = 30\%$， $H_A: p \ne 30\%$

2、计算 u 值

因为 $S_{\hat{p}} = \sqrt{\dfrac{p_0(1-p_0)}{n}} = \sqrt{\dfrac{0.3\times(1-0.3)}{500}} = 0.0205$

于是 $u = \dfrac{\hat{p} - p_0}{S_{\hat{p}}} = \dfrac{0.35 - 0.30}{0.0205} = 2.439$

3、作出统计推断 因为 1.96<u<2.58，0.01<p<0.05，表明样本百分数 $\hat{p}$ =35%与总体百分数 p_0=30%差异显著，该奶牛场的隐性乳房炎比往年严重。

二、两个样本百分数差异显著性检验

在实际工作中，有时需要检验服从二项分布的两个样本百分数差异是否显著。其目的在于检验两个样本百分数 $\hat{p}_1$、$\hat{p}_2$ 所在的两个二项总体百分数 p_1、p_2 是否相同。当两样本的 np、nq 均大于 5 时，可以近似地采用 u 检验法进行检验，但在 np 和（或）nq 小于或等于 30 时，需作连续性矫正。检验的基本步骤是：

（一）提出无效假设与备择假设

$H_0: P_1 = P_2$， $H_A: P_1 \ne P_2$

（二）计算 u 值或 u_c 值

$$u = \frac{\hat{p}_1 - \hat{p}_2}{S_{\hat{p}_1 - \hat{p}_2}} \qquad (5\text{-}11)$$

$$u_c = \frac{|\hat{p}_1 - \hat{p}_2| - 0.5/n_1 - 0.5/n_2}{S_{\hat{p}_1 - \hat{p}_2}} \qquad (5\text{-}12)$$

其中 $\hat{p}_1 = x_1/n_1$， $\hat{p}_2 = x_2/n_2$ 为两个样本百分数，$S_{\hat{p}_1 - \hat{p}_2}$ 为样本百分数差异标准误，计算公式为：

$$S_{\hat{p}_1 - \hat{p}_2} = \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$\bar{p}$ 为合并样本百分数：

$$\bar{p}=\frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}=\frac{x_1+x_2}{n_1+n_2}$$

（三）将 u 或 u_c 的绝对值与 1.96、2.58 比较，作出统计推断 若$|u|$（或$|u_c|$）<1.96, p>0.05，不能否定 $H_0:P_1=P_2$，表明两个样本百分数 $\hat{p}_1$、$\hat{p}_2$ 差异不显著；若 $1.96\le|u|$ (或$|u_c|$)<2.58，0.01<p≤0.05，否定 $H_0:P_1=P_2$，接受 $H_A:P_1\neq P_2$，表明两个样本百分数 $\hat{p}_1$、$\hat{p}_2$ 差异显著；若$|u|$(或$|u_c|$)≥ 2.58，$p\le 0.01$，否定 $H_0:P_1=P_2$，接受 $H_A:P_1\neq P_2$，表明两个样本百分数 $\hat{p}_1$、$\hat{p}_2$ 差异极显著。

【例 5.8】 某养猪场第一年饲养杜长大商品仔猪 9800 头，死亡 980 头；第二年饲养杜长大商品仔猪 10000 头，死亡 950 头，试检验第一年仔猪死亡率与第二年仔猪死亡率是否有显著差异？

此例，两样本死亡率分别为：

$$\hat{p}_1=\frac{x_1}{n_1}=\frac{980}{9800}=10\% \qquad \hat{p}_2=\frac{x_2}{n_2}=\frac{950}{10000}=9.5\%$$

合并的样本死亡率为：

$$\bar{p}=\frac{x_1+x_2}{n_1+n_2}=\frac{980+950}{9800+10000}=9.747\%$$

因为 $n_1\bar{p}=9800\times 9.747\%=955.206$

$n_1\bar{q}=n_1(1-\bar{p})=9800\times(1-9.747\%)=8844.794$

$n_2\bar{p}=10000\times 9.747\%=974$

$n_2\bar{q}=n_2(1-\bar{p})=10000\times(1-9.747\%)=9026$

即 n_1p、n_1q、n_2p、n_2q 均大于 5，并且都大于 30，可利用 u 检验法，不需作连续矫正。检验基本步骤是：

1、提出无效假设与备择假设 $H_0:P_1=P_2$，$H_A:P_1\neq P_2$

2、计算 u 值

因为 $S_{\hat{p}_1-\hat{p}_2}=\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1}+\frac{1}{n_2})}$

$$=\sqrt{9.747\%\times(1-9.747\%)\times(\frac{1}{9800}+\frac{1}{10000})}$$

=0.00422

于是 $u=\frac{\hat{p}_1-\hat{p}_2}{S_{\hat{p}_1-\hat{p}_2}}=\frac{10\%-9.5\%}{0.00422}=1.185$

3、作出统计推断 由于 u<1.96，p>0.05，不能否定 $H_0:P_1=P_2$，表明第一年仔猪死亡率与第二年仔猪死亡率差异不显著。

第四节　总体参数的区间估计

参数估计是统计推断的另一重要内容。所谓参数估计就是用样本统计量来估计总体参数，有点估计（**point estimation**）和区间估计（**interval estimation**）之分。将样本统计量直接作为总体相应参数的估计值叫点估计。点估计只给出了未知参数估计值的大小，没有考虑试验误差的影响，也没有指出估计的可靠程度。区间估计是在一定概率保证下指出总体参数的可能范围，所给出的可能范围叫置信区间（**confidence interval**），给出的概率保证称为置信度或置信概率（**confidence probability**）。本节介绍正态总体平均数 μ 和二项总体百分数 P 的区间估计。

一、正态总体平均数 μ 的置信区间

设有一来自正态总体的样本，包含 n 个观测值 $x_1, x_2, \ldots, x_n$，样本平均数 $\bar{x} = \sum x/n$，标准误 $S_{\bar{x}} = S/\sqrt{n}$。总体平均数为 μ。

因为 $t = (\bar{x} - \mu)/S_{\bar{x}}$ 服从自由度为 n-1 的 t 分布。双侧概率为 a 时，有：

$P(-t_a \le t \le t_a) = 1 - a$，也就是说 t 在区间 $[-t_a, t_a]$ 内取值的可能性为 1-a，即：

$$P(-t_a \le \frac{\bar{x} - \mu}{S_{\bar{x}}} \le t_a) = 1 - a$$

对 $-t_a \le \frac{\bar{x} - \mu}{S_{\bar{x}}} \le t_a$ 变形得：

$$\bar{x} - t_a S_{\bar{x}} \le \mu \le \bar{x} + t_a S_{\bar{x}} \qquad (5\text{-}13)$$

亦即 $$P(\bar{x} - t_a S_{\bar{x}} \le \mu \le \bar{x} + t_a S_{\bar{x}}) = 1 - a$$

（5-13）式称为总体平均数 μ 置信度为 1-a 的置信区间。其中 $t_a S_{\bar{x}}$ 称为置信半径；$\bar{x} - t_a S_{\bar{x}}$ 和 $\bar{x} + t_a S_{\bar{x}}$ 分别称为置信下限和置信上限；置信上、下限之差称为置信距，置信距越小，估计的精确度就越高。

常用的置信度为 95%和 99%，故由（5-13）式可得总体平均数 μ 的 95%和 99%的置信区间如下：

$$\bar{x} - t_{0.05} S_{\bar{x}} \le \mu \le \bar{x} + t_{0.05} S_{\bar{x}} \qquad (5\text{-}14)$$

$$\bar{x} - t_{0.01} S_{\bar{x}} \le \mu \le \bar{x} + t_{0.01} S_{\bar{x}} \qquad (5\text{-}15)$$

【例 5.9】 某品种猪 10 头仔猪的初生重为 1.5、1.2、1.3、1.4、1.8、0.9、1.0、1.1、1.6、1.2（kg），求该品种猪仔猪初生重总体平均数 μ 的置信区间。

经计算得 $\bar{x} = 1.2$，$S_{\bar{x}} = 0.08$，由 $df = n - 1 = 10 - 1 = 9$，查 t 值表得 $t_{0.05(9)} = 2.262$，$t_{0.01(9)} = 3.250$，因此

95%置信半径为 $t_{0.05(df)} S_{\bar{x}} = 2.262 \times 0.08 = 0.18$

95%置信下限为 $\bar{x}-t_{0.05(df)}S_{\bar{x}}=1.2-0.18=1.02$

95%置信上限为 $\bar{x}+t_{0.05(df)}S_{\bar{x}}=1.2+0.18=1.38$

所以该品种仔猪初生重总体平均数 μ 的 95%置信区间为

$$1.02(kg)\le\mu\le1.38(kg)$$

又因为

99%置信半径为 $t_{0.01(df)}S_{\bar{x}}=3.25\times0.08=0.26$

99%置信下限为 $\bar{x}-t_{0.01(df)}S_{\bar{x}}=1.2-0.26=0.94$

99%置信上限为 $\bar{x}+t_{0.01(df)}S_{\bar{x}}=1.2+0.26=1.46$

所以该品种仔猪初生重总体平均数 μ 的 99%置信区间为

$$0.94(kg)\le\mu\le1.46(kg)$$

二、二项总体百分数 P 的置信区间

样本百分数 $\hat{P}$ 只是总体百分数 P 的点估计值。百分数的置信区间则是在一定置信度下对总体百分数作出区间估计。求总体数的置信区间有两种方法：正态近似法和查表法，这里仅介绍正态近似法。

当 $n>1000$，$P\ge1\%$ 时，总体 P 的 95%、99%置信区间为：

$$\hat{P}-1.96S_{\hat{P}}\le P\le\hat{P}+1.96S_{\hat{P}} \qquad (5\text{-}16)$$

$$\hat{P}-2.58S_{\hat{P}}\le P\le\hat{P}+2.58S_{\hat{P}} \qquad (5\text{-}17)$$

其中，$\hat{P}$ 为样本百分数，$S_{\hat{P}}$ 为样本百分数标准误，$S_{\hat{P}}$ 的计算公式为：

$$S_{\hat{P}}=\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \qquad (5\text{-}18)$$

【例 5.10】 调查某地 1500 头奶牛，患结核病的有 150 头，求该地区奶牛结核病患病率的 95%、99%置信区间。

由于 $n=1500>1000$，$\hat{P}=150/1500=10\%>1\%$，采用正态分布近似法求置信区间。

因为

$$S_{\hat{P}}=\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}=\sqrt{\frac{0.1\times(1-0.1)}{1500}}=0.0077$$

所以该地区奶牛结核病患病率 P 的 95%、99%置信区间为：

$$0.1-1.96\times0.0077\le P\le0.1+1.96\times0.0077$$

$$0.1-2.58\times0.0077\le P\le0.1+2.58\times0.0077$$

即

$$8.49\%\le P\le11.15\%$$

$$8.01\%\le P\le11.99\%$$

习　　题

1、为什么在分析试验结果时需要进行显著性检验？检验的目的是什么？

2、什么是统计假设？统计假设有哪几种？各有何含义？

3、显著性检验的基本步骤是什么？什么是显著水平？根据什么确定显著水平？

4、什么是统计推断？为什么统计推断的结论有可能发生错误？有哪两类错误？如何降低两类错误？

5、什么是双侧检验、单侧检验？各在什么条件下应用？二者有何关系？

6、进行显著性检验应注意什么问题？如何理解显著性检验结论中的“差异不显著”、“差异显著”、“差异极显著”？

7、什么是配对试验设计、非配对试验设计？两种设计有何区别？

8、什么是总体平均数 μ 、总体百分数 P 的点估计与区间估计？ μ 与 P 的 95%、99%的置信区间为何？

9、随机抽测了 10 只兔的直肠温度，其数据为：38.7、39.0、38.9、39.6、39.1、39.8、38.5、39.7、39.2、38.4（℃），已知该品种兔直肠温度的总体平均数 μ_0=39.5（℃），试检验该样本平均温度与 μ_0 是否存在显著差异？

（$t=2.641$　$0.01<P<0.05$）

10、11 只 60 日龄的雄鼠在 x 射线照射前后之体重数据见下表（单位：g）：检验雄鼠在照射 X 射线前后体重差异是否显著？

编　号	1	2	3	4	5	6	7	8	9	10	11
照射前	25.7	24.4	21.1	25.2	26.4	23.8	21.5	22.9	23.1	25.1	29.5
照射后	22.5	23.2	20.6	23.4	25.4	20.4	20.6	21.9	22.6	23.5	24.3

（$t=4.132$　$P<0.01$）

11、某猪场从 10 窝大白猪的仔猪中，每窝抽出性别相同、体重接近的仔猪 2 头，将每窝两头仔猪随机地分配到两个饲料组，进行饲料对比试验，试验时间 30 天，增重结果见下表。试检验两种饲料喂饲的仔猪平均增重差异是否显著？

窝号	1	2	3	4	5	6	7	8	9	10
饲料Ⅰ	10.0	11.2	12.1	10.5	11.1	9.8	10.8	12.5	12.0	9.9
饲料Ⅱ	9.5	10.5	11.8	9.5	12.0	8.8	9.7	11.2	11.0	9.0

（$t=3.455$　$P<0.01$）

12、分别测定了 10 只大耳白家兔、11 只青紫蓝家兔在停食 18 小时后正常血糖值如下，问该两个品种家兔的正常血糖值是否有显著差异？（单位：kg）

大耳白	57	120	101	137	119	117	104	73	53	68	
青紫蓝	89	36	82	50	39	32	57	82	96	31	88

（$t=12.455$　$P<0.01$）

13、有人曾对公雏鸡作了性激素效应试验。将 22 只公雏鸡完全随机地分为两组，每组 11 只。一组接受性激素 A（睾丸激素）处理；另一组接受激素 C（雄甾烯醇酮）处理。在第 15 天取它们的鸡冠个别称重，所得数据如下：

激素	鸡冠重量（mg）

A	57	120	101	137	119	117	104	73	53	68	118
C	89	30	82	50	39	22	57	32	96	31	88

问激素 *A* 与激素 *C* 对公雏鸡鸡冠重量的影响差异是否显著。并分别求出接受激素 *A* 与激素 *C* 的公雏鸡鸡冠重总体平均数的 95%、99%置信区间。

（$t=3.376$　*P*<0.01；$77.447\le\mu_1\le116.553$，$69.189\le\mu_1\le124.811$；$(37.301\le\mu_2\le74.699$，$29.404\le\mu_2\le82.596)$

14、某鸡场种蛋常年孵化率为 85%，现有 100 枚种蛋进行孵化，得小鸡 89 只，问该批种蛋的孵化结果与常年孵化率有无显著差异？

（u_c=0. 980　　*P*>0. 05）

15、研究甲、乙两药对某病的治疗效果，甲药治疗病畜 70 例，治愈 53 例；乙药治疗 75 例，治愈 62 例，问两药的治愈率是否有显著差异？并计算两种药物治愈率总体百分率的 95%、99%置信区间。

（u_c=0.829　　*P*>0. 05；$(64.95\%\le P_1\le85.76\%$，$62.48\%\le P_1\le88.94\%)$；$(74.10\%\le P_2\le91.84\%$，$71.34\%\le P_2\le93.94\%)$

第七章　次数资料分析——χ^2检验

前面介绍了计量资料的统计分析方法——t 检验法与方差分析法。在畜牧、水产等科学研究中，除了分析计量资料以外，还常常需要对次数资料、等级资料进行分析。等级资料实际上也是一种次数资料。次数资料服从二项分布或多项分布，其统计分析方法不同于服从正态分布的计量资料。本章将分别介绍对次数资料、等级资料进行统计分析的方法。

第一节　χ^2统计量与χ^2分布

一、χ^2统计量的意义

为了便于理解，现结合一实例说明χ^2（读作卡方）统计量的意义。根据遗传学理论，动物的性别比例是1:1。统计某羊场一年所产的876只羔羊中，有公羔428只，母羔448只。按1:1的性别比例计算，公、母羔均应为438只。以A表示实际观察次数，T表示理论次数，可将上述情况列成表7-1。

表7-1　　羔羊性别实际观察次数与理论次数

性别	实际观察次数A	理论次数T	$A-T$	$(A-T)^2/T$
公	428（A_1）	438（T_1）	-10	0.2283
母	448（A_2）	438（T_2）	10	0.2283
合计	876	876	0	0.4566

从表7-1看到，实际观察次数与理论次数存在一定的差异，这里公、母各相差10只。 这个差异是属于抽样误差(把对该羊场一年所生羔羊的性别统计当作是一次抽样调查)、还是羔羊性别比例发生了实质性的变化?要回答这个问题，首先需要确定一个统计量用以表示实际观察次数与理论次数偏离的程度；然后判断这一偏离程度是否属于抽样误差，即进行显著性检验。为了度量实际观察次数与理论次数偏离的程度，最简单的办法是求出实际观察次数与理论次数的差数。从表7-1看出：$A_1-T_1=-10$，$A_2-T_2=10$，由于这两个差数之和为0， 显然不能用这两个差数之和来表示实际观察次数与理论次数的偏离程度。为了避免正、负抵消，可将两个差数A_1-T_1、A_2-T_2平方后再相加，即计算$\Sigma(A-T)^2$，其值越大，实际观察次数与理论次数相差亦越大，反之则越小。但利用$\Sigma(A-T)^2$表示实际观察次数与理论次数的偏离程度尚有不足。例如某一组实际观察次数为505、理论次数为500，相差5；而另一组实际观察次数为26、 理论次数为21，相差亦为5。显然这两组实际观察次数与理论次数的偏离程度是不同的。因为前者是相对于理论次数500相差5，后者是相对于理论次数21相差5。为了弥补这一不足，可先将各差数平方除以相应的理论次数后再相加，并记之为χ^2，即

$$\chi^2=\sum\frac{(A-T)^2}{T} \tag{7-1}$$

也就是说χ^2是度量实际观察次数与理论次数偏离程度的一个统计量，χ^2越小，表明实

际观察次数与理论次数越接近；χ^2=0，表示两者完全吻合；χ^2越大，表示两者相差越大。

对于表7-1的资料，可计算得

$$\chi^2=\sum\frac{(A-T)^2}{T}=\frac{(-10)^2}{438}+\frac{10^2}{438}=0.4566$$

表明实际观察次数与理论次数是比较接近的。

二、χ^2分布

上面在属于离散型随机变量的次数资料的基础上引入了统计量χ^2， 它近似地服从统计学中一种连续型随机变量的概率分布——χ^2分布。下面对统计学中的χ^2分布作一简略介绍。

设有一平均数为μ、方差为σ^2的正态总体。现从此总体中独立随机抽取n个随机变量：x_1、x_2、…、x_n，并求出其标准正态离差：

$$u_1=\frac{x_1-\mu}{\sigma},\quad u_2=\frac{x_2-\mu}{\sigma},\ldots,\quad u_n=\frac{x_n-\mu}{\sigma}$$

记这n个相互独立的标准正态离差的平方和为χ^2：

$$\chi^2=u_1^2+u_2^2+\ldots+u_n^2=\sum u_i^2=\sum(\frac{x_i-\mu}{\sigma})^2=\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}\qquad(7\text{-}2)$$

它服从自由度为n的χ^2分布，记为

$$\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}\sim\chi^2_{(n)};$$

若用样本平均数$\bar{x}$代替总体平均数μ，则随机变量

$$\chi^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2}=\frac{(n-1)S^2}{\sigma^2}\qquad(7\text{-}3)$$

服从自由度为n-1的χ^2分布，记为

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2_{(n-1)}$$

因此，χ^2分布是由正态总体随机抽样得来的一种连续型随机变量的分布。显然，$\chi^2\geqslant0$，即χ^2的取值范围是[0, +∞)；χ^2分布密度曲线是随自由度不同而改变的一组曲线。随自由度的增大，曲线由偏斜渐趋于对称；$df\geqslant30$时，$\sqrt{2\chi^2}$ 接近平均数为$\sqrt{2df-1}$ 的正态分布。

图7-1给出了几个不同自由度的χ^2概率分布密度曲线。

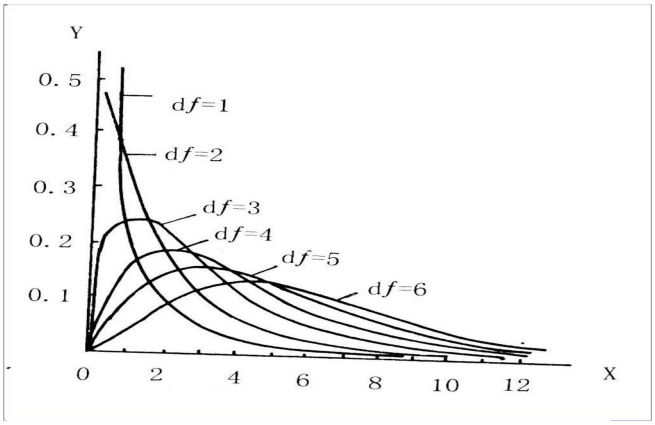


图 7-1　几个自由度的χ^2概率分布密度曲

三、χ^2的连续性矫正

由(7-1)式计算的χ^2只是近似地服从连续型随机变量χ^2分布。在对次数资料进行χ^2检验利用连续型随机变量χ^2分布计算概率时，常常偏低，特别是当自由度为1时偏差较大。Yates(1934)提出了一个矫正公式，矫正后的χ^2值记为χ_c^2：

$$\chi_c^2=\sum\frac{(|A-T|-0.5)^2}{T} \tag{7-4}$$

当自由度大于1时，(7-1)式的χ^2分布与连续型随机变量χ^2分布相近似，这时，可不作连续性矫正，但要求各组内的理论次数不小于5。若某组的理论次数小于5，则应把它与其相邻的一组或几组合并，直到理论次数大于5为止。

第二节　适合性检验

一、适合性检验的意义

判断实际观察的属性类别分配是否符合已知属性类别分配理论或学说的假设检验称为适合性检验。在适合性检验中，无效假设为H_0：实际观察的属性类别分配符合已知属性类别分配的理论或学说；备择假设为H_A：实际观察的属性类别分配不符合已知属性类别分配的理论或学说。并在无效假设成立的条件下，按已知属性类别分配的理论或学说计算各属性类别的理论次数。因所计算得的各个属性类别理论次数的总和应等于各个属性类别实际观察次数的总和，即独立的理论次数的个数等于属性类别分类数减1。也就是说，适合性检验的自由度等于属性类别分类数减1。若属性类别分类数为k，则适合性检验的自由度为$k-1$。然后根据(7-1)或(7-4)计算出χ^2或χ_c^2。将所计算得的χ^2或χ_c^2值与根据自由度$k-1$查χ^2值表(附表8)所得的临界χ^2值：$\chi^2_{0.05}$、$\chi^2_{0.01}$比较：若χ^2（或χ_c^2）$<\chi^2_{0.05}$，$P>0.05$，表明实际观察次数

与理论次数差异不显著，可以认为实际观察的属性类别分配符合已知属性类别分配的理论或学说；若$\chi^2_{0.05}\leqslant\chi^2$（或$\chi^2_c$）$<\chi^2_{0.01}$，0.01＜$P$≤0.05，表明实际观察次数与理论次数差异显著，实际观察的属性类别分配不符合已知属性类别分配的理论或学说；　若χ^2（或χ^2_c）$\geqslant\chi^2_{0.01}$，P≤0.01，表明实际观察次数与理论次数差异极显著，实际观察的属性类别分配极显著不符合已知属性类别分配的理论或学说。

二、适合性检验的方法

下面结合实例说明适合性检验方法。

【例 7.1】 在进行山羊群体遗传检测时，观察了 260 只白色羊与黑色羊杂交的子二代毛色，其中 181 只为白色，79 只为黑色，问此毛色的比率是否符合孟德尔遗传分离定律的 3∶1 比例？

检验步骤如下：

（一）提出无效假设与备择假设

H_0：子二代分离现象符合 3∶1 的理论比例。

H_A：子二代分离现象不符合 3∶1 的理论比例。

（二）选择计算公式　由于本例是涉及到两组毛色（白色与黑色），属性类别分类数 k=2，自由度 df=k-1=2-1=1，须使用公式（7—4）来计算 χ^2_c。

（三）计算理论次数　根据理论比率 3∶1 求理论次数：

白色理论次数：T_1=260×3/4=195

黑色理论次数：T_2=260×1/4=65

或 T_2=260-T_1=260-195=65

（四）计算 χ^2_c

表 7—2　　χ^2_c 计算表

性　状	实际观察次数（A）	理论次数（T）	A-T	χ^2_c
白　色	181	195	−14	0.935
黑　色	79	65	+14	2.804
总　和	260	260	0	3.739

$$\chi^2_C=\Sigma\frac{(|A-T|-0.5)^2}{T}=\frac{(|181-195|-0.5)^2}{195}+\frac{(|79-65|-0.5)^2}{65}=3.739$$

（五）查临界 χ^2 值，作出统计推断　当自由度 df=1 时，查得 $\chi^2_{0.05(1)}$ =3.84，计算的 $\chi^2_c<\chi^2_{0.05(1)}$，故 P>0.05，不能否定 H_0，表明实际观察次数与理论次数差异不显著，可以认为白色羊与黑色羊的比率符合孟德尔遗传分离定律 3∶1 的理论比例。

【例 7.2】 在研究牛的毛色和角的有无两对相对性状分离现象时，用黑色无角牛和红色有角牛杂交，子二代出现黑色无角牛 192 头，黑色有角牛 78 头，红色无角牛 72 头，红色有角牛 18 头，共 360 头。试问这两对性状是否符合孟德尔遗传规律中 9∶3∶3∶1 的遗传比例？

检验步骤：

（一）提出无效假设与备择假设

H_0：实际观察次数之比符合 9∶3∶3∶1 的理论比例。

H_A：实际观察次数之比不符合 9∶3∶3∶1 的分离理论比例。

（二）选择计算公式 由于本例的属性类别分类数 k=4：自由度 $df=k$-1=4-1=3>1，故利用（7—1）式计算 χ^2。

（三）计算理论次数 依据各理论比率 9:3:3:1 计算理论次数：

黑色无角牛的理论次数 T_1：360×9/16=202.5；

黑色有角牛的理论次数 T_2：360×3/16=67.5；

红色无角牛的理论次数 T_3：360×3/16=67.5；

红色有角牛的理论次数 T_4：360×1/16=22.5。

或 T_4=360-202.5-67.5-67.5=22.5

（四）列表计算 χ^2

表 7—3 χ^2 计算表

类 型	实际观察次数 A	理论次数 T	A-T	$(A-T)^2/T$
黑色无角牛	192（A_1）	202.5（T_1）	−10.5	0.5444
黑色有角牛	78（A_2）	67.5（T_2）	+10.5	1.6333
红色无角牛	72（A_3）	67.5（T_3）	+4.5	1.6333
红色有角牛	18（A_4）	22.5（T_4）	−4.5	0.9000
总 计	360	360	0	4.711

$$\chi^2=\sum\frac{(A-T)^2}{T}=0.5444+1.6333+1.6333+0.9=4.711$$

（五）查临界 χ^2 值，作出统计推断 当 df=3 时，$\chi^2_{0.05(3)}$=7.815，因 $\chi^2<\chi^2_{0.05(3)}$，P>0.05，不能否定 H_0，表明实际观察次数与理论次数差异不显著，可以认为毛色与角的有无两对性状杂交二代的分离现象符合孟德尔遗传规律中 9∶3∶3∶1 的遗传比例。

*三、χ^2 显著性检验的再分割法

当实际观察次数与理论次数经 χ^2 检验差异显著或极显著时，还应对其结果进行再分割检验，下面举例说明。

【例 7.3】 两对相对性状杂交子二代 4 种表现型 *A-B-*、*A-bb*、*aaB-*、*aabb* 的观察次数依次为 152、39、53、6，问这两对相对性状的遗传是否符合孟德尔遗传规律中 9∶3∶3∶1 的比例。

检验步骤同**【例 7.2】**，计算结果见表 7—4。

表 7—4 χ^2 计算表

表现型	实际观察次数 A	理论次数 T	A-T	$(A-T)^2/T$
A-B-	152	140.625	11.375	0.920
A-bb	39	46.875	-7.875	1.323

aa B-	53	46.875	6.125	0.800
aa bb	6	15.625	-9.625	5.929
总　和	250	250	0	χ^2 =8.972

表中理论次数依 9∶3∶3∶1 理论比率计算：

*A-B-*的理论次数 T_1：250×9/16=140.625；

A-bb 的理论次数 T_2=*aaB-*的理论次数 T_3：250×3/16=46.875；

aa bb 的理论次数 T_4：250×1/16=15.625。

或 T_4=250-140.625-46.875-46.875=15.625

由表 7—5 可知χ^2=8.972，由 *df*=3 查χ^2 值表得：$\chi^2_{0.05(3)}$=7.815, $\chi^2_{0.01(3)}$=11.345。因为 $\chi^2_{0.05(3)}$< χ^2< $\chi^2_{0.01(3)}$，故 0.01<*P*<0.05,表明实际观察次数与理论观察次数差异显著，即该资料不符合 9∶3∶3∶1 的遗传规律，有必要进一步检验，以具体确定哪样的表现型的实际观察次数不符合 9∶3∶3∶1 的比例。这时须采用χ^2 检验的再分割法。

χ^2 检验的再分割法的具体作法是：将一张列联表的总χ^2 统计量，分割为数目等于该表总自由度的多个分量。每个分量的χ^2 值对应于由原始数据所产生的一特殊列联表，且每个分量独立于其它分量，这样各分量的χ^2 值之和等于总χ^2 值。这种可加性只有在所分割的列联表是相互独立、各分量的χ^2 值不作矫正的条件下成立。

下面我们利用χ^2 检验的再分割法对**【例 7. 3】**的资料进行进一步检验。

1. 检验 *A-B-*，*A-bb*, *aaB-* 3 种表现型是否符合 9∶3∶3 的比例。分割后χ^2 值（记为 χ_1^2）的计算见表 7—5。

表 7—5　χ_1^2 计算表（理论比例 9∶3∶3）

表现型	实际观察次数 *A*	理论次数 *T*	*A-T*	*(A-T)* ^{2}*/T*
A-B-	152	146.400	5.600	0.214
A-bb	39	48.800	-9.800	1.968
aaB-	53	48.800	4.200	0.361
总和	244	244	0	2.543

χ_1^2 =0.214+1.968+0.361=2.543

由 df_1=3-1=2，查χ^2 值表得，$\chi^2_{0..05(2)}$=5.991， χ_1^2 <$\chi^2_{0..05}$，*P*>0.05，表明实际观察次数与理论观察次数差异不显著，可以认为 3 种表现型符合 9∶3∶3 的理论比例。于是，我们再分析表现型 *aabb* 是否与其它三种表现型的合并组比例不符合 1:15 的理论比例。

2. 检验 *aabb* 表现型与其它三种表现型的合并组是否符合 1:15 的比例，分割后χ^2 值（记为 χ_2^2）的计算见表 7-6。

表 7—6　χ_2^2 分割表（理论比例 1∶15）

表现型	实际观察次数 *A*	理论次数 *T*	*A-T*	*(A-T)* ^{2}*/T*
aabb	6	15.625	-9.625	5.929
其它三种表现型合并组	244	234.375	9.625	0.395
总　和	250	250.000	0	6.324

χ_2^2=5.929+0.395=6.324

由 df_2=2-1=1，查χ^2 表得，$\chi^2_{0.05(12)}$=3.841，$\chi^2_{0.05(1)}$=6.635，由于$\chi^2_{0.05(1)}<\chi^2_2<\chi^2_{0.01(1)}$，故 0.01<*P*<0.05，表明实际观察次数与理论次数差异显著，即 *aabb* 表现型与其它三种表现型组合不符合 1∶15 的比例，这样的结论可为我们进一步研究这个问题提供线索。

χ^2 经分割后，χ_1^2=2.543，χ_2^2=6.324，$\chi_1^2+\chi_2^2$=8.867 与总χ^2=8.922 略有差异，这是由于计算的舍入误差所造成；总自由度 *df*=3，df_1=2，df_2=1，所以总 $df=df_1+df_2$。如果分割后χ^2 值或自由度之和不等于χ^2 值或总自由度，说明所分割的列联表相互不独立。

*四、资料分布类型的适合性检验

实际观测得来的资料是否服从某种理论分布，亦可应用适合性检验来判断。在正态分布的适合性检验中，由于理论次数是由样本总次数、平均数与标准差决定的，用去 3 个自由度，所以自由度为 *k*-3（*k* 为组数）；而在二项分布和波松分布的适合性检验中，由于其理论次数由总次数与均数求得，丧失 2 个自由度，所以自由度为 *k*-2。但应注意，当组段内理论次数小于 5 时，必须与相邻组段进行合并，直至合并的理论次数大于 5 时为止。下面分别举例说明。

（一）实际观测资料服从正态分布的适合性检验

【例 7. 4】 检验 200 头大白猪仔猪一月窝重的资料是否服从正态分布。

表 7—7 200 头大白猪仔猪一月龄窝重服从正态分布的适合性检验表

组限 (1)	组中值(*x*) (2)	实际次数 (*f*) (3)	上限（*l*） (4)	$l-\bar{x}$ （5）	$u=(x-\bar{x})/S_c$ (6)	累加概率（*a*） (7)	各组概率 (8)	理论次数 (9)	χ^2 (10)
<8		0	8	-57.6	-2.57	0.0051	0.0051	1.016	
8—	12	4 } 10	16	-49.6	-2.21	0.0136	0.0085	1.704 } 6.44	1.9680
16—	20	6	24	-41.6	-1.85	0.0322	0.0186	3.720	
24—	28	9	32	-33.6	-1.50	0.0668	0.0346	6.920	0.6252
32—	36	10	40	-25.6	-1.14	0.1271	0.0603	12.060	0.3519
40—	44	13	48	-17.6	-0.78	0.2177	0.0906	18.120	1.4467
48—	52	17	56	-9.6	-0.43	0.3336	0.1159	23.180	1.6476
56—	60	26	64	-1.6	-0.07	0.4721	0.1385	27.700	0.1043
64—	68	35	72	6.4	0.29	0.6141	0.1420	28.400	1.5338
72—	76	28	80	14.4	0.64	0.7389	0.1248	24.960	0.3703
80—	84	21	88	22.4	1.00	0.8413	0.1024	20.480	0.0132
88—	92	16	96	30.4	1.35	0.9115	0.0702	14.040	0.2736
96—	100	8	104	38.4	1.71	0.9564	0.0449	8.980	0.1069
104–	108	4	112	46.4	2.07	0.9808	0.0244	4.880	
112–	116	3 } 7	120	54.4	2.42	0.99224	0.0114	2.288 } 8.72	0.3393
>120		0					0.0078	1.552	
合计			200				1.0000	200.00	8.7308

1、先将资料（原始数据略）整理成次数分布表，组限、组中值、各组的次数列于表 7-7 的（1）、（2）、（3）栏，再将各组上限列于第（4）栏中。

2、计算各组组上限与均数（$\bar{x}$ =65.6kg）之差，列于第（5）栏。

3、计算校正标准差 S_c。由于由分组资料求得的标准差较不分组时所得标准差为大，故需作校正。

$$S_c=\sqrt{\frac{\sum fx^2-\frac{(\sum fx)^2}{n}}{n-1}-\frac{i^2}{12}}=\sqrt{\frac{961920-\frac{(13120)^2}{200}}{200-1}-\frac{8^2}{12}}=22.44(kg)$$

4、依公式$u=\frac{x-\bar{x}}{S_c}$求各组上限的正态离差，列入第 6 栏。如第一组

$$u=\frac{8-65.6}{22.44}=-2.57$$

5、设该资料服从正态分布，依 u 值查正态分布表得各组段的累计概率（a），列入第（7）栏。如当 u=-2.57 时，a=0.0051，u=0.29 时，a=0.6141。

6、求出每一组段的概率，列入第（8）栏。由下一组段的累加概率减去本组段的累加概率而得。如 8— 组段的概率为 0.0136-0.0051=0.0085。

7、以总数 n=200 头乘以各组概率便得理论次数，列入第（9）栏。凡理论值小于 5 者应加以合并。本例前三组与后三组分别合并。合并后的实际次数与理论次数分别为 10 与 6.44、7 与 8.72，见第(3)与第（9）栏。

8、求各组χ^2值，列入第（10）栏。

9、确定自由度。这里是因为求理论次数时用去均数，标准差与总次数三个统计量，该例经合并共 12 个组，故 df=12-3=9。

10、结论。由 df=9 查χ^2表得：$\chi^2_{0.05(9)}$=16.919，而计算所得的χ^2值为：χ^2=8.7808，因为$\chi^2<\chi^2_{0.05}$， P>0.05，表明各组实际次数与由正态分布计算的理论次数差异不显著，可以认为大白猪仔猪一月窝重服从正态分布。

（二）实际观测资料服从二项分布的适合性检验

【例 7.5】 用 800 粒牧草种子进行发芽试验，分 80 行，每行 10 粒种子，共有 174 粒发芽。则每粒种子发芽的概率约为 174/800=0.2175，不发芽的概率约为 0.7825（即 1-0.2175），每行发芽种子数见表 7—8，问该资料是否服从二项分布。

表 7—8　80 行发芽试验资料服从二项分布的适合性检验计算表

一行内种子发芽数	实际行数 A	理论概率	理论行数 T	$\chi^2=(A-T)^2/T$
0	6	0.0861	6.8880	0.1145
1	20	0.2392	19.1360	0.0390
2	28	0.2992	23.9360	0.6900
3	12	0.2218	17.7440	1.8594
4	8	0.1079	8.6320	
5	6	0.0360	2.8800	
6	0	0.0083	0.6640	
7	0	0.0013	0.1040	0.2176
8	0	0.0001	0.0800	
9	0	0.0000	0.0000	
10	0	0.0000	0.0000	
	14		12.3600	
总　和	80			2.9205

表中理论概率由二项分布概率计算公式：$C_n^k p^k q^{n-k}$ 计算，如

$$C_{10}^0 p^0 q^{10}=\frac{10!}{10!0!}\times 0.2175^0\times 0.7825^{10}=0.0861;$$

$$C_{10}^{1}p^{1}q^{9}=\frac{10!}{9!1!}\times 0.2175^{1}\times 0.7825^{9}=0.2392\ ;$$

表中的理论行数由理论概率乘以 80 行而得，如

$$0.0861\times 80=6.8880,$$

$$0.2392\times 80=19.1360$$

由于表中后 6 组的理论次数均小于 5，故将后 6 组与第 5 组合并为一组。并组以后，资料分为 5 组。

由表 7—8 可知，χ^2=2.9025。由 df=5-2=3，查χ^2 值表得：$\chi^2_{0.05(3)}$=7.81，因为$\chi^2<\chi^2_{0.05}$，P>0.05，表明实际行数与由二项分布计算得来的理论行数差异不显著，可以认为种子发芽试验的结果服从二项分布。

（三）实际观测资料服从波松分布的适合性检验

【例 7.6】 用显微镜检查某样品内结核菌的数目，对某些视野内各小方格的结核菌数计数，然后按不同的结核菌数目把格子分类，记录每类的格子数。其结果见表 7—9 第（1）、（2）栏。试检验结核菌数是否服从波松分布。

1. 计算理论概率　设结核菌数服从波松分布 P（λ），其概率计算公式为：

$$P_m=\frac{\lambda^m}{m!}e^{-\lambda}\quad(\lambda>0)\qquad(7—5)$$

其中 λ 为平均数 μ，且等于方差 σ^2。此时因 λ 未知，可利用样本平均数 $\bar{x}$ 来估计。利用加权法求样本平均数 $\bar{x}$ 为：

$$\bar{x}=\frac{\Sigma fx}{\Sigma f}=\frac{5\times 0+19\times 1+\cdots+1\times 9}{118}=2.983$$

表 7—9　　结核菌数服从波松分布适合性检验计算表

结核菌数 x（1）	实际格子数(f=A) (2)	理论概率（3）	理论格子数(T)(4)	(A-T)^2/T(5)
0	5	0.0506		
1	19	0.1511	5.9708	0.1578
2	26	0.2253	17.8298	0.1768
3	26	0.2240	26.5854	0.0129
4	21	0.1671	26.4320	0.0071
5	13	0.0997	19.7178	0.0834
6	5 ⎫	0.0496	5.8528 ⎫	
7	1 ⎬ 8	0.0211	2.4898 ⎬ 9.5818	0.1297
8	1 ⎪	0.0079	0.9322 ⎪	0.2611
9	1 ⎭	0.0026	0.3068 ⎭	
总计	118	0.9990	117.8820	0.7288

将 $\bar{x}$ 代入（7—5）式，求得各项理论概率为

$$P(x=k)=\frac{2.983^k}{k!}e^{-2.983},m=0,1,2,\cdots,9$$

计算结果列于第（3）栏。

2. 计算理论次数　将总次数 N=118 乘以各组的理论概率即得各组理论次数 T。计算结果列于第（4）栏。由于表后 4 组的理论次数小于 5，故将后 4 组与第 7 组合并为一组，

合并后的实际格子数为 8，理论格子数为 9.5818。

3. 计算χ^2值　根据表 7—9 第（5）栏的数据可得χ^2值为：

$$\chi^2 = \Sigma\frac{(A-T)^2}{T} = 0.7288$$

因为此例经并组后的分组数为 7；计算理论次数利用了样本平均数和总次数，所以自由度为 7-2=5。当 df=5 时，查χ^2值表得：$\chi^2_{0.05(5)}$=11.07，因为$\chi^2 < \chi^2_{0.05(5)}$，$P$>0.05，表明结核菌的各实际格子数与根据波松分布计算出的理论格子数差异不显著，可以认为结核菌数服从波松分布。

第三节　独立性检验

一、独立性检验的意义

对次数资料，除进行适合性检验外，有时需要分析两类因子是相互独立还是彼此相关。如研究两类药物对家畜某种疾病治疗效果的好坏，先将病畜分为两组，一组用第一种药物治疗，另一组用第二种药物治疗，然后统计每种药物的治愈头数和未治愈头数。这时需要分析药物种类与疗效是否相关，若两者彼此相关，表明疗效因药物不同而异，即两种药物疗效不相同；若两者相互独立，表明两种药物疗效相同。这种根据次数资料判断两类因子彼此相关或相互独立的假设检验就是独立性检验。独立性检验实际上是基于次数资料对子因子间相关性的研究。

独立性检验与适合性检验是两种不同的检验方法，除了研究目的不同外，还有以下区别：

（一）独立性检验的次数资料是按两因子属性类别进行归组。根据两因子属性类别数的不同而构成 2×2、2×c、r×c 列联表（r 为行因子的属性类别数，c 为列因子的属性类别数）。而适合性检验只按某一因子的属性类别将如性别、表现型等次数资料归组。

（二）适合性检验按已知的属性分类理论或学说计算理论次数。独立性检验在计算理论次数时没有现成的理论或学说可资利用，理论次数是在两因子相互独立的假设下进行计算。

（三）在适合性检验中确定自由度时，只有一个约束条件：各理论次数之和等于各实际次数之和，自由度为属性类别数减 1。而在 r×c 列联表的独立性检验中，共有 rc 个理论次数，但受到以下条件的约束：1、rc 个理论次数的总和等于 rc 个实际次数的总和；2、r 个横行中的每一个横行理论次数总和等于该行实际次数的总和。但由于 r 个横行实际次数之和的总和应等于 rc 个实际次数之和，因而独立的行约束条件只有 r-1 个；3、类似地，独立的列约束条件有 c-1 个。因而在进行独立性检验时，自由度为 rc-1-(r-1)-(c-1)=(r-1)(c-1)，即等于（横行属性类别数-1）×（直列属性类别数-1）。

二、独立性检验的方法

下面结合实例分别介绍 2×2、2×c、r×c 列联表独立性检验的具体过程。

（一）2×2 列联表的独立性检验　2×2 列联表的一般形式如表 7—10 所示，其自由度 df=(C -1) (r-1)=(2-1) (2-1)=1，在进行χ^2检验时，需作连续性矫正，应计算 χ^2_c 值。

表 7—10　　2×2 列联表的一般形式

	1	2	行总合$T_{i.}$
1	A_{11}（T_{11}）	A_{12}（T_{12}）	$T_{1.}=A_{11}+A_{12}$
2	A_{21}（T_{21}）	A_{22}（T_{22}）	$T_{2.}=A_{21}+A_{22}$
列总合$T_{.j}$	$T_{.1}=A_{11}+A_{21}$	$T_{.2}=A_{12}+A_{22}$	$T_{..}=A_{11}+A_{12}+A_{21}+A_{22}$

其中 A_{ij} 为实际观察次数，T_{ij} 为理论次数。

【例 7.7】 某猪场用 80 头猪检验某种疫苗是否有预防效果。结果是注射疫苗的 44 头中有 12 头发病，32 头未发病；未注射的 36 头中有 22 头发病，14 头未发病，问该疫苗是否有预防效果？

1、 先将资料整理成列联表（见表 7—11）

表 7—11　　2×2 列联表

	发病	未发病	行总和 $T_{i.}$	发病率
注射	12(18.7)	32(25.3)	$T_{1.}$：44	27.3%
未注射	22(15.3)	14(20.7)	$T_{2.}$：36	61.1%
列总和 $T_{.j}$	$T_{.1}$：34	$T_{.2}$：46	$T_{..}$：80	

2、 提出无效假设与备择假设

H_0：发病与否和注射疫苗无关，即二因子相互独立。

H_A：发病与否和注射疫苗有关，即二因子彼此相关。

3、 计算理论次数　根据二因子相互独立的假设，由样本数据计算出各个理论次数。二因子相互独立，就是说注射疫苗与否不影响发病率。也就是说注射组与未注射组的理论发病率应当相同，均应等于总发病率 34/80=0.425。依此计算出各个理论次数如下：

注射组的理论发病数：T_{11}=44×34/80=18.7

注射组的理论未发病数：T_{12}=44×46/80=25.3，　或：T_{12}=44-18.7=25.3；

未注射组的理论发病数：T_{21}=36×34/80=15.3，或 T_{21}=34-18.7=15.3；

未注射组的理论未发病数：T_{22}=36×46/80=20.7，或 T_{22}=36-15.3=20.7。

从上述各理论次数 T_{ij} 的计算可以看到，理论次数的计算利用了行、列总和，总总和，4 个理论次数仅有一个是独立的。表 7-11 括号内的数据为相应的理论次数。

4、 计算 χ_c^2 值　将表 7-11 中的实际次数、理论次数代入 7—4 式得：

$$\chi_c^2=\frac{(|12-18.7|-0.5)^2}{18.7}+\frac{(|32-25.3|-0.5)^2}{25.3}+\frac{(|22-15.3|-0.5)^2}{15.3}$$

$$+\frac{(|14-20.7|-0.5)^2}{20.7}=7.944$$

5、 由自由度 df=1 查临界 χ^2 值，作出统计推断　因为 $\chi^2_{0.01(1)}$=6.63，而 χ_c^2=7.944>$\chi^2_{0.01(1)}$，P<0.01，否定 H_0，接受 H_A，表明发病率与是否注射疫苗极显著相关，这里表现为注射组发病率极显著低于未注射组，说明该疫苗是有预防效果的。

在进行 2×2 列联表独立性检验时，还可利用下述简化公式（7-6）计算 χ_c^2：

$$\chi_c^2=\frac{(|A_{11}A_{22}-A_{12}A_{21}|-T../2)^2T..}{T._1T._2T_1.T_2.} \qquad (7—6)$$

在（7-6）式中，不需要先计算理论次数，直接利用实际观察次数 A_{ij}，行、列总和 $T_{i.}$、$T_{.j}$ 和总总和 $T..$进行计算，比利用公式（7-4）计算简便，且舍入误差小。

对于【例 7. 7】，利用（7-6）式可得：

$$\chi_c^2=\frac{(|12\times14-32\times22|-\frac{80}{2})^2\times80}{34\times46\times36\times44}=7.944$$

所得结果与前面计算计算的相同。

（二）2×c 列联表的独立性检验　2×c 列联表是行因子的属性类别数为 2，列因子的属性类别数为 c（c≥3）的列联表。其自由度 df=(2-1) (c -1)，因为 c≥3，所以自由度大于 2，在进行χ^2 检验时，不需作连续性矫正。2×c 表的一般形式见表 7—12。

表 7—12　　2×c 联列表一般形式

	1	2	…	c	行总和$T_{i.}$
1	A_{11}	A_{12}	…	A_{1C}	$T_{1.}$
2	A_{21}	A_{22}	…	A_{2c}	$T_{2.}$
列总和$T_{.j}$	$T_{.1}$	$T_{.2}$	…	$T_{.c}$	总总和$T..$

其中 A_{ij}（i=1，2；j=1，2，…，c）为实际观察次数。

【例 7. 8】 在甲、乙两地进行水牛体型调查，将体型按优、良、中、劣四个等级分类，其结果见表 7—13，问两地水牛体型构成比是否相同。

表 7—13　　两地水牛体型分类统计

	优	良	中	劣	行总和 $T_{i.}$
甲	10（13.3）	10(10.0)	60(53.3)	10(13.4)	90
乙	10(6.7)	5(5.0)	20(26.7)	10(6.6)	45
列总和 $T_{.j}$	20	15	80	20	135

这是一个 2×4 列联表独立性检验的问题。检验步骤如下：

1．提出无效假设与备择假设

H_0：水牛体型构成比与地区无关，即两地水牛体型构成比相同。

H_A：水牛体型构成比与地区有关，即两地水牛体型构成比不同。

2．计算各个理论次数，并填在各观察次数后的括号中　计算方法与 2×2 表类似，即根据两地水牛体型构成比相同的假设计算。如优等组中，甲地、乙地的理论次数按理论比率 20/135 计算；良等组中甲地、乙地的理论次数按理论比率 15/135 计算；中等、劣等组中甲地、乙地的理论次数分别按理论比率 80/135 和 20/135 计算。

甲地优等组理论次数：T_{11}=90×20/135=13.3，

乙地优等组理论次数：T_{21}=45×20/135=6.7，或 T_{21}=20-13.3=6.7；

其余各个理论次数的计算类似。

3. 计算计算χ^2 值

$$\chi^2=\frac{(10-13.3)^2}{13.3}+\frac{(10-10)^2}{10}+\cdots\cdots+\frac{(20-26.7)^2}{26.7}+\frac{(10-6.6)^2}{6.6}=7.582$$

4. 由自由度 df=3 查临界χ^2 值，作出统计推断　因为$\chi^2_{0..05(3)}$=7.815，而χ^2=7.582<$\chi^2_{0..05(3)}$，p>0.05，不能否定 H_0,可以认为甲、乙两地水牛体型构成比相同。

在进行 2×c 列联表独立性检验时，还可利用下述简化公式（7-7）或（7-8）计算 χ^2：

$$\chi^2=\frac{T_{..}^{\ 2}}{T_1.T_2.}[\Sigma\frac{A_{1j}^2}{T_{.j}}-\frac{T_1^2.}{T_{..}}] \qquad (7—7)$$

或

$$\chi^2=\frac{T_{..}^{\ 2}}{T_1.T_2.}[\Sigma\frac{A_{2j}^2}{T_{.j}}-\frac{T_2^2.}{T_{..}}] \qquad (7—8)$$

（7-7）或（7-8）式的区别在于：（7-7）式利用第一行中的实际观察次数 A_{1j} 和行总和 $T_{1.}$；（7-8）式利用第二行中的实际观察次数 A_{2j} 和行总和 $T_{2.}$，计算结果相同。对于[例 7-7]利用（7-8）式计算χ^2 值得：

$$\chi^2=\frac{135^2}{90\times45}[\frac{10^2}{20}+\frac{5^2}{15}+\frac{20^2}{80}+\frac{10^2}{20}-\frac{45^2}{135}]=7.502$$

计算结果与利用（7—1）式计算的结果因舍入误差略有不同。

此外，在畜牧、水产科学研究中，有时需将数量性状资料以等级分类，如剪毛量分为特等、一等、二等，产奶量分为高产与低产等，这些由数量性状资料转化为质量性状的次数资料检验，也可用χ^2 检验。

【例 7.9】 分别统计了 *A*、*B* 两个品种各 67 头经产母猪的产仔情况，结果见表 7—14，问 *A*、*B* 两品种的产仔构成比是否相同？

表 7—14　　*A*、*B* 两个品种产仔数的分类统计

	9 头以下	10-12 头	13 头以上	行总和 $T_{i.}$
A	17	44	6	67
B	5	33	29	67
列总和 $T._j$	22	77	35	134

1、提出无效假设与备择假设

H_0：*A*、*B* 两个品种产仔数分级构成比相同。

H_A：*A*、*B* 两个品种产仔数分级构成比不同。

2、计算χ^2 值　　用简化公式（7—7）计算为：

$$\chi^2=\frac{134^2}{67\times67}+[\frac{17^2}{22}+\frac{44^2}{77}+\frac{6^2}{35}-\frac{67^2}{134}]=23.23$$

3、由自由度 df=(2-1)(3-1)=2 查临界χ^2 值，作出统计推断　因为$\chi^2_{0.05(2)}$=9.21，χ^2> $\chi^2_{0.01}$，　P<0.01，所以否定 H_0，接受 H_A ，表明 *A*、*B* 两品种产仔数构成比差异极显著。需要应用χ^2 检验的再分割法来具体确定分级构成比差异在那样的等级。

4、 χ^2 检验的再分割法

（1）先对两个品种产仔数在 9 头以下和 10—12 头进行χ^2 检验，分割后的情况见表 7

—15。

表 7—15　χ^2_1 计算表

	9 头以下	10-12 头	行总和 $T_{i\cdot}$
A	17	44	61
B	5	33	38
列总和 $T_{\cdot j}$	22	77	99

利用简化公式（7-7）计算χ^2_1值为：

$$\chi_1^2 = \frac{99^2}{61\times 38}[\frac{17^2}{22}+\frac{44^2}{77}-\frac{61^2}{99}] = 2.930$$

由 df_1=2-1=1，查χ^2值表得：$\chi^2_{0.05(1)}$=3.841，因为$\chi^2_1 < \chi^2_{0.05(1)}$，$P$>0.05，表明这两个品种的产仔数在 9 头以下和 10—12 头这两个级别内的比率差异不显著。

（2）对产仔数在 13 头以上组与其他合并组（即 9 头以下和 10—12 头两个组的合并）进行χ^2检验，分割后见表 7—16。

表 7—16　χ^2_2 计算表

	合并组	13 头以上	总　和
A	61	6	67
B	38	29	67
总　和	99	35	134

利用简化公式（7-7）计算χ^2_2值为：

$$\chi_2^2 = \frac{134^2}{67\times 67}[\frac{61^2}{99}+\frac{6^2}{35}-\frac{67^2}{134}] = 20.458$$

由 df_2=2-1=1，查χ^2值表得：$\chi^2_{0.05(1)}$=3.846，$\chi^2_{0.01(1)}$=6.63，因为$\chi^2_2 > \chi^2_{0.01(1)}$，$P$<0.01，表明这两个品种的产仔数在合并组与 13 头以上组的比率差异极显著。其中 *B* 品种产仔数在 13 头以上的比率为 29/67=42.38%，极显著高于 *A* 品种产仔数在 13 头以上的比率 6/67=8.96%。或者说 *B* 品种产仔数在合并组（12 头以下）的比率为 38/67=56.72%，极显著低于 *A* 品种产仔数在合并组（12 头以下）的比率 61/67=91.04%。

经分割检验后，$df=df_1+df_2$=2+1=3, χ^2=23.25=χ^2_1+ χ^2_2=2.93+20.458=23.388，χ^2 略小于$\chi^2_1+\chi^2_2$，是由于计算中的舍入误差所致。

（三）$r\times c$ 列联表的独立性检验　$r\times c$ 表是指行因子的属性类别数为 r（r>2），列因子的属性类别数为 c(c>2)的列联表。其一般形式见表 7-17。

表 7—17　$r\times c$ 列联表的一般形式

	1	2	…	c	行总和$T_{i\cdot}$
1	A_{11}	A_{12}	…	A_{1C}	$T_{1\cdot}$
2	A_{21}	A_{22}	…	A_{2c}	$T_{2\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮
r	A_{r1}	A_{r2}	…	A_{rc}	$T_{r\cdot}$
列总和$T_{\cdot j}$	$T_{\cdot 1}$	$T_{\cdot 2}$	…	$T_{\cdot c}$	$T_{\cdot\cdot}$

其中 A_{ij}（i=1, 2, …r; j=1, 2, …c）为实际观察次数。

$r\times c$ 列联表各个理论次数的计算方法与上述（2×2）、（2×c）表适合性检验类似。但一般用简化公式计算χ^2值，其公式为：

$$\chi^2 = T..[\Sigma\frac{A_{ij}^2}{T_i.T._j}-1] \qquad (7—9)$$

【例 7.10】 对三组奶牛（每组 39 头）分别喂给不同的饲料，各组发病次数统计如下表，问发病次数的构成比与所喂饲料是否有关？

表 7—18　三组牛的发病次数资料

发病次数	饲　料			总　和
	1	2	3	
0	19（17.3）	16(17.3)	17(17.3)	52
1	1(0.3)	0(0.3)	0(0.3)	1
2	0(1.3)	3(1.3)	1(1.3)	4
3	7(5.7)	9(5.7)	1(5.7)	17
4	3(4.7)	5(4.7)	6(4.7)	14
5	4(3.3)	1(3.3)	5(3.3)	10
6	2(2.0)	1(2.0)	3(2.0)	6
7	0(1.3)	2(1.3)	2(1.3)	4
8	1(2.3)	2(2.3)	4(2.3)	7
9	2(0.7)	0(0.7)	0(0.7)	2
总　和	39	39	39	117

检验步骤如下：

1、提出无效假设与备择假设

H_0：发病次数的构成比与饲料种类无关，即二者相互独立。

H_A：发病次数的构成比与饲料种类有关，即二者彼此独立。

2、计算理论次数　对于理论次数小于 5 者，将相邻几个组加以合并（见表 7—19），合并后的各组的理论次数均大于 5。

表 7—19　资料合并结果

发病次数	饲　料			
	1	2	3	总　和
0	19(17.3)	16(17.3)	17(17.3)	52
1-3	8(7.3)	12(7.3)	2(7.3)	22
4-5	7(8.0)	6(8.0)	11(8.0)	24
6-8	5(6.3)	5(6.3)	9(6.3)	19
总　和	39	39	39	117

（注：括号内为理论次数）

3、计算χ^2值　利用（7-9）式计算χ^2值，得：

$$\chi^2 = 117[\frac{19^2}{39\times52}+\frac{16^2}{39\times52}+\cdots\cdots+\frac{5^2}{39\times19}+\frac{9^2}{39\times19}-1] = 10.61$$

4、查临界χ^2值，进行统计推断　由自由度 df=(4-1)(3-1)=6，查临界χ^2值得：$\chi^2_{0..05(6)}$=12.9，因为计算所得的χ^2< $\chi^2_{0.05(6)}$，P>0.05，不能否定 H_O，可以认为奶牛的发病次数的构成比与饲料种类相互独立，即用三种不同的饲料饲喂奶牛，各组奶牛发病次数的构成比相同。

习　　题

1. χ^2 检验与 t 检验、F 检验在应用上有什么区别？

2. 什么是适合性检验和独立性检验？它们有何区别？

3. 什么情况下χ^2检验需作矫正？如何矫正？什么情况下先将各组合并后再作χ^2检验？合并时应注意什么问题？

4. 在什么情况下需应用χ^2 检验的再分割法？如何对总χ^2 值进行分割？

5. 两对相对性状杂交子二代 A—B—，A—bb，aaB—，$aabb$ 4 种表现型的观察次数依次为：315、108、101、32，问是否符合 9∶3∶3∶1 的遗传比例？（x^2=0.475，接受 H_0）

6. 某猪场 102 头仔猪中，公的 54 头，母的 48 头，问是否符合家畜性别 1∶1 的理论比例。

（χ^2=0.2450，p>0.05）

7. 某生物药品厂研制出一批新的鸡瘟疫苗，为检验其免疫力，用 200 只鸡进行试验，某中注射 100 只（经注射后患病的 10 只，不患病的 90 只），对照组（注射原疫苗组）100 只（经注射后患病的 15 只，不患病的 85 只），试问新旧疫苗的免疫力是否有差异。

（χ^2=0.731，接受 H_0）

8. 甲、乙、丙三个奶牛场高产奶牛、低产奶牛头数统计如下，试问三个奶牛场高、低产奶牛的构成比是否有差异。

场　地	高产奶牛	低产奶牛
甲	32	18
乙	28	26
丙	38	10

（χ^2=8.269，0.01<p<0.05，需进一步作χ^2 分割检验）。

9. 某防疫站对屠宰场及食品零售点的猪肉进表皮沙门氏杆菌代菌情况进行检验，结果如下表，问屠宰场与零售点猪肉带菌率有无显著差异。

采样地点	带菌头数	不带菌头数
屠宰场	8	32
零售点	14	16

（χ^2_c=4.486，P<0.05）

10. 对陕西三个秦川牛保种基地县进行秦川牛肉用性能外形调查，划分为优良中下 4 个等级，试问三个地区秦川牛肉用性能各级构成比差异是否显著。

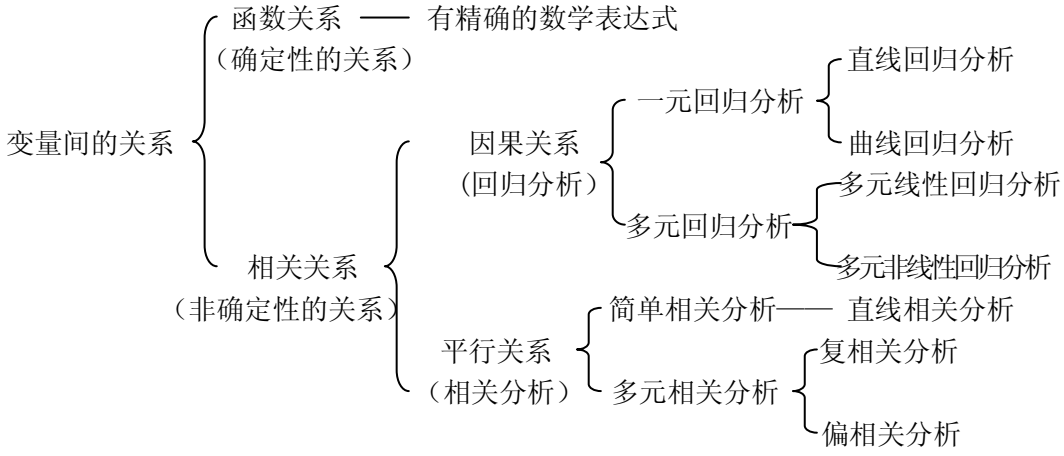
地区	优	良	中	下
甲	10	10	60	10
乙	10	5	20	10
丙	5	5	23	6

（χ^2=7.73，P>0.05）

第八章　直线回归与相关

前面各章我们讨论的问题，都只涉及到一个变量，如体重、日增重或发病率。但是，由于客观事物在发展过程中相互联系、相互影响，因而在畜牧、水产等试验研究中常常要研究两个或两个以上变量间的关系。变量间的关系有两类，一类是变量间存在着完全确定性的关系，可以用精确的数学表达式来表示，如长方形的面积（S）与长（a）和宽（b）的关系可以表达为：$S=ab$。它们之间的关系是确定性的，只要知道了其中两个变量的值就可以精确地计算出另一个变量的值，这类变量间的关系称为函数关系。另一类是变量间关系不存在完全的确定性关系，不能用精确的数学公式来表示，如人的身高与体重的关系；仔猪初生重与断奶重的关系；猪瘦肉率与背膘厚度、眼肌面积、胴体长等的关系等等，这些变量间都存在着十分密切的关系，但不能由一个或几个变量的值精确地求出另一个变量的值。像这样一类关系在生物界中是大量存在的，统计学中把这些变量间的关系称为相关关系，把存在相关关系的变量称为相关变量。

相关变量间的关系一般分为两种，一种是因果关系，即一个变量的变化受另一个或几个变量的影响，如仔猪的生长速度受遗传、营养、饲养管理等因素的影响，子女的身高受父母身高的影响；另一种是平行关系，即两个以上变量之间共同受到另外因素的影响，如人的身高和体重之间的关系，兄弟身高之间的关系等都属于平行关系。变量间的关系及分析方法归纳如下：



统计学上采用回归分析（**regression analysis**）研究呈因果关系的相关变量间的关系。表示原因的变量称为自变量，表示结果的变量称为依变量。研究“一因一果”，即一个自变量与一个依变量的回归分析称为一元回归分析；研究“多因一果”，即多个自变量与一个依变量的回归分析称为多元回归分析。一元回归分析又分为直线回归分析与曲线回归分析两种；多元回归分析又分为多元线性回归分析与多元非线性回归分析两种。回归分析的任务是揭示出呈因果关系的相关变量间的联系形式，建立它们之间的回归方程，利用所建立的回归方程，由自变量（原因）来预测、控制依变量（结果）。

统计学上采用相关分析（**correlation analysis**）研究呈平行关系的相关变量之间的关系。对两个变量间的直线关系进行相关分析称为简单相关分析（也叫直线相关分析）；对多个变量进行相关分析时，研究一个变量与多个变量间的线性相关称为复相关分析；研究其余变量保

持不变的情况下两个变量间的线性相关称为偏相关分析。在相关分析中，变量无自变量和依变量之分。相关分析只能研究两个变量之间相关的程度和性质或一个变量与多个变量之间相关的程度，不能用一个或多个变量去预测、控制另一个变量的变化，这是回归分析与相关分析区别的关键所在。但是二者也不能截然分开，因为由回归分析可以获得相关的一些重要信息，由相关分析也能获得回归的一些重要信息。

本章先介绍直线回归与相关分析。

第一节　直线回归

一、直线回归方程的建立

对于两个相关变量，一个变量用符号 x 表示，另一个变量用 y 表示，如果通过试验或调查获得两个变量的成对观测值，可表示为（x_1，y_1），（x_2，y_2），……，（x_n，y_n）。为了直观地看出 x 和 y 间的变化趋势，可将每一对观测值在平面直角坐标系描点，作出散点图（见图 8-1）。

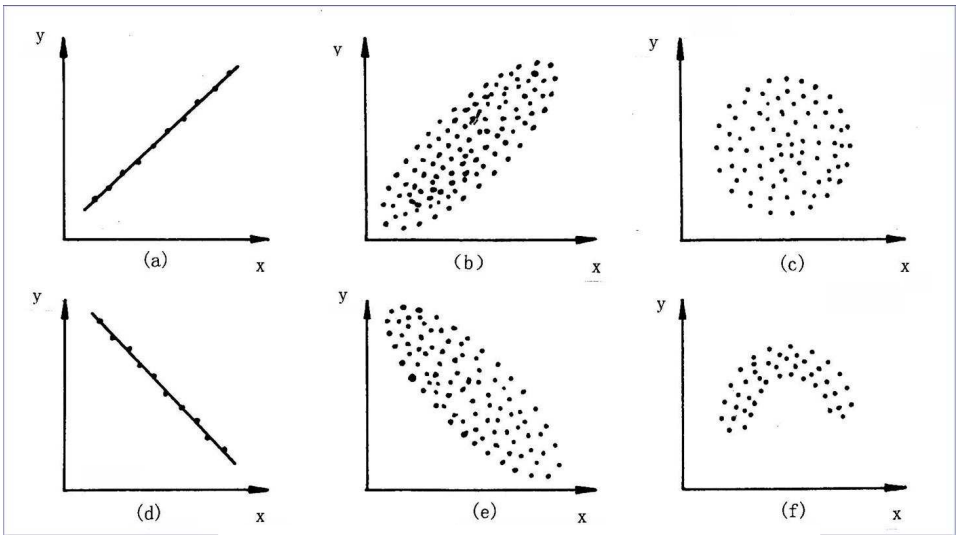


图 8-1 （x，y）的散点图

从散点图（图 8-1）可以看出：①两个变量间关系的性质（是正相关还是负相关）和程度（是相关密切还是不密切）；②两个变量间关系的类型，是直线型还是曲线型；③是否有异常观测值的干扰。散点图直观地、定性地表示了两个变量之间的关系。为了探讨它们之间的规律性，还必须根据观测值将其内在关系定量地表达出来。

如果两个相关变量间的关系是直线关系，根据 n 对观测值所描出的散点图，如图 8—1（c）和图 8—1（d）。如果把变量 y 与 x 内在联系的总体直线回归方程记为 $y=\alpha+\beta x$，由于依变量的实际观测值总是带有随机误差，因而实际观测值 y_i 可表示为：

$$y_i = x_i + \beta x_i + \varepsilon_i \qquad (i=1, 2, \cdots, n) \qquad (8—1)$$

其中 ε_i 为相互独立，且都服从 N（0，σ^2）的随机变量。这就是直线回归的数学模型。我们

可以根据实际观测值对α，β以及方差σ^2做出估计。

在x,y的直角坐标平面上可以作出无数条直线，而回归直线是指所有直线中最接近散点图中全部散点的直线。设样本直线回归方程为：

$$\hat{y}=a+bx \tag{8-2}$$

其中，a是α的估计值，b是β的估计值。

回归直线在平面坐标系中的位置取决于a、b的取值，为了使$\hat{y}=a+bx$能最好地反应y和x两变量间的数量关系，根据最小二乘法，a、b应使回归估计值与观测值的偏差平方和最小，即：

$Q=\sum(y-\hat{y})^2=\sum(y-a-bx)^2=$最小。

根据微积分学中的极值原理，令Q对a、b的一阶偏导数等于0，即：

$$\frac{\partial Q}{\partial a}=-2\sum(y-a-bx)=0$$

$$\frac{\partial Q}{\partial b}=-2\sum(y-a-bx)x=0$$

整理得关于a、b的正规方程组：

$$\begin{cases} an+b\sum x=\sum y \\ a\sum x+b\sum x^2=\sum xy \end{cases}$$

解正规方程组，得：

$$b=\frac{\sum xy-(\sum x)(\sum y)/n}{\sum x^2-(\sum x)^2/n}=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}=\frac{SP_{xy}}{SS_x} \tag{8-3}$$

$$a=\bar{y}-b\bar{x} \tag{8-4}$$

（8-3）式中的分子是自变量x的离均差与依变量y的离均差的乘积和$\sum(x-\bar{x})(y-\bar{y})$，简称乘积和，记作$SP_{xy}$，分母是自变量$x$的离均差平方和$\sum(x-\bar{x})^2$，记作$SS_x$。

a叫做样本回归截距，是回归直线与y轴交点的纵坐标，当x=0时，$\hat{y}$=a；b叫做样本回归系数，表示x改变一个单位，y平均改变的数量；b的符号反映了x影响y的性质，b的绝对值大小反映了x影响y的程度。

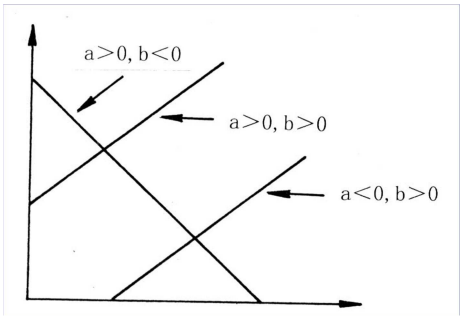


图 8-2 直线回归方程$\hat{y}=a+bx$的图象

a和b均可取正值，也可取负值，因具体资料而异，由图8-2可以看出，a>0，表示回

归直线在第一象限与 y 轴相交；$a<0$ 表示回归直线在第一象限与 x 轴相交。$b>0$，表示 y 随 x 的增加而增加；$b<0$；表示 y 随 x 的减少而减少；$b=0$ 或与 0 差异不显著时，表示 y 的变化与 x 的取值无关，两变量间不存在直线回归关系。这只是对 a 和 b 的统计学解释，对于具体资料，a 和 b 往往还有专业上的实际意义。

$\hat{y}$ 叫做回归估计值，是当 x 在在其研究范围内取某一个值时，y 值平均数 $\alpha+\beta x$ 估计值。研究 y 和 $\hat{y}$ 间的关系，可发现回归方程的三个基本性质：

性质 1　$Q=\sum(y-\hat{y})^2=$最小；

性质 2　$\sum(y-\hat{y})=0$；

性质 3　回归直线必须通过中心点 $(\bar{x},\bar{y})$。

如果将（8-3）式代入（8-2）式，得到回归方程的另一种形式：

$$\hat{y}=\bar{y}-b\bar{x}+bx=\bar{y}+b(x-\bar{x}) \qquad (8\text{-}5)$$

【例 8.1】在四川白鹅的生产性能研究中，得到如下一组关于雏鹅重（g）与 70 日龄重(g)的数据，试建立 70 日龄重(y)与雏鹅重(x)的直线回归方程。

表 8-1 四川白鹅重与 70 日龄重测定结果　（单位：g）

编号	1	2	3	4	5	6	7	8	9	10	11	12
雏鹅重(x)	80	86	98	90	120	102	95	83	113	105	110	100
70 日龄重(y)	2350	2400	2720	2500	3150	2680	2630	2400	3080	2920	2960	2860

1、作散点图　以雏鹅重（x）为横坐标，70 日龄重（y）为纵坐标作散点图，见图 8-3。由图形可见四川白鹅的 70 日龄重与雏鹅重间存在直线关系，70 日龄重随雏鹅重的增大而增大。

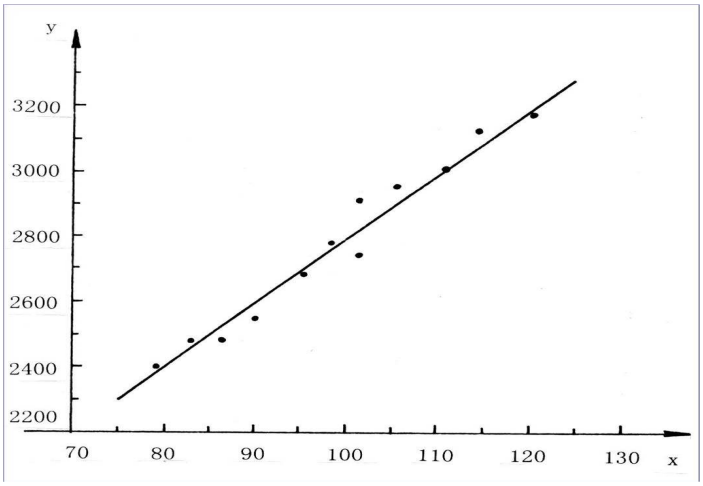


图 8-3 四川白鹅的雏鹅重与 70 日龄重散点图和回归直线图

2、计算回归截距 a，回归系数 b，建立直线回归方程

首先根据实际观测值计算出下列数据：

$$\bar{x}=\sum x/n=1182/12=98.5$$

$$\bar{y}=\sum y/n=32650/12=2720.8333$$

$$SS_x=\sum x^2-(\sum x)^2/n=118112-(1182)^2/12=1685.00$$

$$SP_{xy}=\sum xy-\frac{(\sum x)(\sum y)}{n}=3252610-\frac{1182\times 32650}{12}=36585.00$$

$$SS_y=\sum y^2-(\sum y)^2/n=89666700-(32650)^2/12=831491.67$$

进而计算出 b、a：

$$b=\frac{SP_{xy}}{SS_x}=\frac{36585}{1685.00}=21.7122$$

$$a=\bar{y}-b\bar{x}=2720.8333-21.7122\times 98.5=582.1816$$

得到四川白鹅的 70 日龄重 y 对雏鹅重 x 的直线回归方程为：

$$\hat{y}=582.1816+21.7122x$$

从回归系数可知，雏鹅重每增加 1g，70 日龄平均重增加 21.7122g。

根据直线回归方程可作出回归直线，见图 8-3。从图 8-3 可看出，尽管 $\hat{y}=582.1816+21.7122x$ 是该资料最恰当的回归方程，但是并不是所有的散点都恰好落在回归直线上，这说明用 $\hat{y}$ 去估计 y 是有偏差的。

3、**直线回归的偏离度估计**　以上根据使偏差平方和 $\sum(y-\hat{y})^2$ 最小建立了直线回归方程。偏差平方和 $\sum(y-\hat{y})^2$ 的大小表示了实测点与回归直线偏离的程度，因而偏差平方和又称为离回归平方和。统计学已经证明：在直线回归分析中离回归平方和的自由度为 n-2。于是可求得离回归均方为：$\sum(y-\hat{y})^2/(n-2)$。离回归均方是模型（8-1）中 σ^2 的估计值。离回归均方的平方根叫离回归标准误，记为 S_{yx}，即

$$S_{yx}=\sqrt{\sum(y-\hat{y})^2/(n-2)} \qquad (8\text{-}6)$$

离回归标准误 S_{yx} 的大小表示了回归直线与实测点偏差的程度，即回归估测值 $\hat{y}$ 与实际观测值 y 偏差的程度，于是我们把离回归标准误 S_{yx} 用来表示回归方程的偏离度。离回归标准误 S_{yx} 大表示回归方程偏离度大，S_{yx} 小表示回归方程偏离度小。

在用（8-6）式计算离回归标准误时，需要把每一个 x 值的回归估计值 $\hat{y}$ 计算出来，因而计算麻烦，且累计舍入误差大。以后我们将证明：

$$\sum(y-\hat{y})^2=SS_y-SP_{xy}^2/SS_x \qquad (8\text{-}7)$$

利用（8-7）式先计算出 $\sum(y-\hat{y})^2$，然后再代入（8-6）式求 S_{yx}，这样就简便多了。对于【例 8.1】有

$$\sum(y-\hat{y})^2=SS_y-SP_{xy}^2/SS_x=83149167-36585^2/1685=37152.07$$

所以

$$S_{yx}=\sqrt{\sum(y-\hat{y})^2/(n-2)}=\sqrt{37152.07/(12-2)}=60.9525\ (g)$$

即当利用直线回归 $\hat{y}=582.1816+21.7122x$，由四川白鹅的雏鹅重估计 70 日龄重时，离回归标准误为 60.9525g。

二、 直线回归的显著性检验

若 x 和 y 变量间并不存在直线关系，但由 n 对观测值（x_i，y_i）也可以根据上面介绍的方法求得一个回归方程 $\hat{y}=a+bx$。显然，这样的回归方程所反应的两个变量间的直线关系是不真实的。这取决于如何判断直线回归方程所反应的两个变量间的直线关系的真实性呢？这取决于变量 x 与 y 间是否存在直线关系。我们先探讨依变量 y 的变异，然后再作出统计推断。

1、直线回归的变异来源

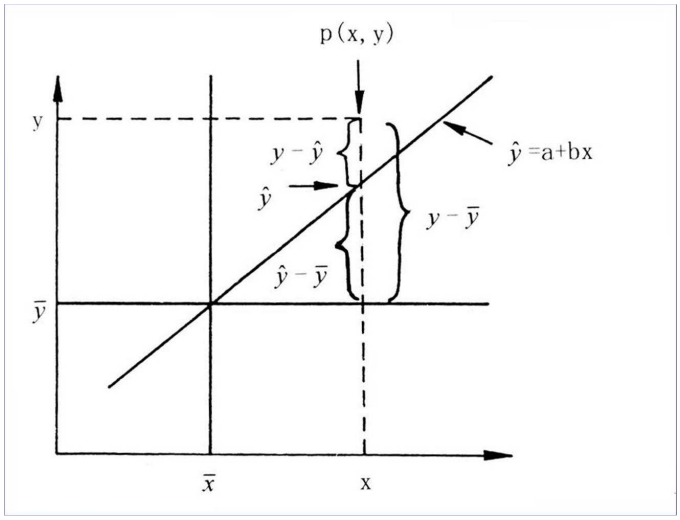


图 8-4 $(y-\bar{y})$ 的分解图

从图8-4看到：依变量y的总变异 $(y-\bar{y})$ 由 y 与 x 间存在直线关系所引起的变异 $(\hat{y}-\bar{y})$ 与偏差 $(y-\hat{y})$ 两部分构成，即

$$(y-\bar{y})=(\hat{y}-\bar{y})+(y-\hat{y})$$

上式两端平方，然后对所有的 n 点求和，则有

$$\sum(y-\bar{y})^2=\sum[(\hat{y}-\bar{y})+(y-\hat{y})]^2$$
$$=\sum(\hat{y}-\bar{y})^2+\sum(y-\hat{y})^2+2\sum(\hat{y}-\bar{y})(y-\hat{y})$$

由于 $\hat{y}=a+bx=\bar{y}+b(x-\bar{x})$，所以 $\hat{y}-\bar{y}=b(x-\bar{x})$

于是
$$\sum(\hat{y}-\bar{y})(y-\hat{y})=\sum b(x-\bar{x})(y-\hat{y})$$
$$=\sum b(x-\bar{x})[(y-\bar{y})-b(x-\bar{x})]$$
$$=\sum b(x-\bar{x})(y-\bar{y})-\sum b(x-\bar{x})\cdot b(x-\bar{x})$$
$$=b\cdot SP_{xy}-b^2\cdot SS_x$$
$$=\frac{SP_{xy}}{SS_x}\cdot SP_{xy}-\left(\frac{SP_{xy}}{SS_x}\right)^2\cdot SS_x=0$$

所以有
$$\sum(y-\bar{y})^2=\sum(\hat{y}-\bar{y})^2+\sum(y-\hat{y})^2 \qquad (8\text{-}8)$$

$\sum(y-\bar{y})^2$ 反映了 y 的总变异程度，称为 y 的总平方和，记为 SS_y；$\sum(\hat{y}-\bar{y})^2$ 反映了

由于 y 与 x 间存在直线关系所引起的 y 的变异程度，称为回归平方和，记为 SS_R；$\sum(y-\hat{y})^2$ 反映了除 y 与 x 存在直线关系以外的原因，包括随机误差所引起的 y 的变异程度，称为离回归平方和或剩余平方和，记为 SS_r。（8-8）式又可表示为：

$$SS_y = SS_R + SS_r \tag{8-9}$$

这表明 y 的总平方和划分为回归平方和与离回归平方和两部分。与此相对应，y 的总自由度 df_y 也划分为回归自由度 df_R 与离回归自由度 df_r 两部分，即

$$df_y = df_R + df_r \tag{8-10}$$

在直线回归分析中，回归自由度等于自变量的个数，即 $df_R = 1$；y 的总自由度 $df_y = n-1$；离回归自由度 $df_r = n-2$。于是：

离回归均方 $MS_r = SS_r / df_r$，回归均方 $MS_R = SS_R / df_R$

2、回归关系显著性检验—F 检验

x 与 y 两个变量间是否存在直线关系，可用 F 检验法进行检验。由（8-1）式可推知，若 x 与 y 间不存在直线关系，则总体回归系数β=0，若 x 与 y 间存在直线关系，则总体回归系数β≠0。所以，对 x 与 y 间是否存在直线关系的假设检验其无效假设 H_0：β=0，备择假设 H_A：β≠0。在无效假设成立的条件下，回归均方与离回归均方的比值服从 $df_1 = 1$ 和 $df_2 = n-2$ 的 F 分布，所以可以用

$$F = \frac{MS_R}{MS_r} = \frac{MS_R / df_R}{SS_r / df_r} = \frac{SS_R}{SS_r /(n-2)} \quad ，\ df_1=1,\ df_2=n-2 \tag{8-11}$$

来检验回归关系即回归方程的显著性。

回归平方和还可用下面的公式计算得到：

$$SS_R = \sum(\hat{y}-\bar{y})^2 = \sum[b(x-\bar{x})]^2$$

$$= b^2\sum(x-\bar{x})^2 = b^2 SS_x = bSP_{xy} \tag{8-12}$$

$$= \frac{SP_{xy}}{SS_x}\cdot SP_{xy} = \frac{SP_{xy}^2}{SS_x} \tag{8-13}$$

利用（8-13）式计算 SS_R 的舍入误差最小；而（8-12）式便于推广到多元线性回归分析的情况。根据（8-9）式。可得到离回归平方和计算公式为：

$$SS_r = SS_y - SS_R = SS_y - \frac{SP_{xy}^2}{SS_x}$$

对于【例 8.1】资料，有

$$SS_y = 831491.67，\qquad SP_{xy} = 36585.00，\qquad SS_x = 1685.00$$

$$SS_R = \frac{SP_{xy}^2}{SS_x} = \frac{36585.00^2}{1685.00} = 794339.60$$

$$SS_r = SS_y - SS_R = 831491.67 - 794339.60 = 37152.07$$

而 $df_y = n-1 = 12-1 = 11, df_R = 1, df_r = 12-2 = 10$。于是可以列出方差分析表进行回归关系显著性检验。

表 8-2　四川白鹅 70 日龄重与雏鹅重回归关系方差分析

变异来源	*df*	*SS*	*MS*	*F* 值	$F_{0.05}$	$F_{0.01}$
回归	1	794339.60	794339.60	213.81**	4.96	10.04
离回归	10	37152.07	3715.21			
总变异	11	831491.67				

因为 $F=213.81>F_{0.01(1,10)}=10.04, P<0.01$，表明四川白鹅 70 日龄重与雏鹅重间存在显著的直线关系。

3、回归系数的显著性检验—*t* 检验

采用回归系数的显著性检验—*t* 检验也可检验 *x* 与 *y* 间是否存在直线关系。回归系数显著性检验的无效假设和备择假设分别为 H_0：β＝0，H_A：β≠0。

t 检验的计算公式为：

$$t=\frac{b}{S_b},\ df=n-2 \qquad (8\text{-}14)$$

$$S_b=\frac{S_{yx}}{\sqrt{SS_x}} \qquad (8\text{-}15)$$

其中，S_b 为回归系数标准误。

对于【例 8.1】资料，已计算得 $SS_x=1685.00, S_{yx}=60.9525$，故有

$$S_b=S_{yx}/\sqrt{SS_x}=60.9525/\sqrt{1685}=1.4849$$

$$t=\frac{b}{S_b}=\frac{21.7122}{1.4849}=14.62$$

当 $df=n-2=12-2=10$, 查 *t* 值表，得 $t_{0.05(10)}=2.228$，$t_{0.01(10)}=3.169$。因 $t=14.62>t_{0.01(10)}$，$P<0.01$，否定 H_0：β＝0，接受 H_A：β≠0，即四川白鹅 70 日龄重（*y*）与雏鹅重（*x*）的直线回归系数 *b*=21.7122 是极显著的，表明四川白鹅 70 日龄重与雏鹅重间存在极显著的直线关系，可用所建立的直线回归方程来进行预测和控制。

F 检验的结果与 *t* 检验的结果一致。事实上，统计学已证明，在直线回归分析中，这二种检验方法是等价的，可任选一种进行检验。

由于四川白鹅 70 日龄重与雏鹅重间的直线回归关系极显著，因此，在实际生产中，可以通过四川白鹅的雏鹅重对 70 日龄重作出预测或控制。特别要指出的是：利用直线回归方程进行预测或控制时，一般只适用于原来研究的范围，不能随意把范围扩大，因为在研究的范围内两变量是直线关系，这并不能保证在这研究范围之外仍然是直线关系。若需要扩大预测和控制范围，则要有充分的理论依据或进一步的实验依据。利用直线回归方程进行预测或控制，一般只能内插，不要轻易外延。

*三、直线回归的区间估计

前面已求出了总体回归截距 *a*、回归系数 *β* 和 *x* 所对应的 *y* 值总体平均数 *a*+*βx* 的估计值 *a*，*b* 和 $\hat{y}$。这仅是一种点估计。下面在一定置信度下对 *α*、*β* 以及 *α*+*βx* 作出区间估计。

1、总体回归截距 a 的置信区间　统计学已证明 $\frac{a-\alpha}{S_a}$ 服从自由度为 n-2 的 t 分布。

其中，S_a 叫做样本回归截距标准误，计算公式为：

$$S_a = S_{yx}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{SS_x}}$$

容易导出 α 的 95%、99%置信区间为：

$$[a - t_{0.05(n-2)}S_a, a + t_{0.05(n-2)}S_a]$$

$$[a - t_{0.01(n-2)}S_a, a + t_{0.01(n-2)}S_a]$$

【例 8.2】　试计算**【例 8.1】**资料回归截距 α 的 95%和 99%置信区间。

对于**【例 8.1】**资料，因为

$$a = 582.1816,\quad S_{yx} = 60.9525,\quad n = 12,\quad \bar{x} = 98.5,\quad SS_x = 1685.00$$

所以，

$$S_a = S_{yx}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{SS_x}} = 60.9525\times\sqrt{\frac{1}{12}+\frac{98.50^2}{1685.00}} = 147.3153$$

$$t_{0.05(n-2)} = t_{0.05(10)} = 2.228,\quad t_{0.01(n-2)} = t_{0.01(10)} = 3.169$$

于是总体回归截距 α 的 95%和 99%置信区间分别为：

[582.1816-2.228×147.3153, 582.1816+2.228×147.3153]

[582.1816-3.169×147.3153, 582.1816+3.169×147.3153]

即[253.9631, 910.40]和[115.3394, 1049.0238]。

这说明在研究雏鹅重与 70 日龄重的关系时，总体回归截距 α 在[253.9631, 910.40]区间内，其可靠度为 95%；在[115.3394, 1049.0238]区间内，其可靠度为 99%。

2、总体回归系数 β 的置信区间　统计学已证明 $\frac{b-\beta}{S_b}$ 服从自由度为 n-2 的 t 公布，

其中，S_b 叫做样本回归系数标准误，由（8-15）式计算。可以导出 β 的 95%、99%置信区间为：

$$[b - t_{0.05(n-2)}S_b, b + t_{0.05(n-2)}S_b] \tag{8-16}$$

$$[b - t_{0.01(n-2)}S_b, b + t_{0.01(n-2)}S_b] \tag{8-17}$$

【例 8.3】　求出**【例 8.1】**资料总体回归系数 β 的 95%和 99%置信区间。

对于**【例 8.1】**资料，因为

$$b = 21.7122,\quad S_{yx} = 60.9525,\quad SS_x = 1685.00$$

$$S_b = S_{yx}/\sqrt{SS_x} = 60.9525/\sqrt{1685} = 1.4849$$

$$t_{0.05(n-2)} = t_{0.05(10)} = 2.228,\quad t_{0.01(n-2)} = t_{0.01(10)} = 3.169$$

所以总体回归系数 β 的 95%和 99%置信区间分别为：

[21.7122-2.228×1.4849, 21.7122+2.228×1.4849]

[21.7122-3.169×1.4849, 21.7122+3.169×1.4849]

即[18.4038, 25.0206]和[17.0066, 26.4178]。

这说明雏鹅重和 70 日龄重的总体回归系数 β 在[18.4038, 25.0206]区间内，其可靠度

为 95%；在[17.0066, 26.4178]区间内，其可靠度为 99%。

3、总体平均数 $\alpha+\beta x$ 的置信区间　统计学已证明 $\frac{\hat{y}-(\alpha+\beta x)}{S_{\hat{y}}}$ 服从自由度为 $n-2$ 的 t 分布。其中 $S_{\hat{y}}$ 叫回归估计标准误，计算公式为：

$$S_{\hat{y}}=S_{yx}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{SS_x}} \tag{8-18}$$

于是可以导出 $\alpha+\beta x$ 的 95%、99%置信区间为：

$$[\hat{y}-t_{0.05(n-2)}S_{\hat{y}},\hat{y}+t_{0.05(n-2)}S_{\hat{y}}] \tag{8-19}$$

$$[\hat{y}-t_{0.01(n-2)}S_{\hat{y}},\hat{y}+t_{0.01(n-2)}S_{\hat{y}}] \tag{8-20}$$

【例 8.4】 求出**【例 8.1】**资料当 x=98 时 y 总体平均数 $\alpha+\beta x$ 的 95%和 99%置信区间。

对于**【例 8.1】**资料，当 x=98 时，$\hat{y}$=582.1816+21.7122×98=2709.9772，而

$$S_{\hat{y}}=S_{yx}\cdot\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{SS_x}}=60.9525\times\sqrt{\frac{1}{12}+\frac{(98-98.5)^2}{1685}}=17.6111$$

$t_{0.05(n-2)}=t_{0.05(10)}=2.228$，$t_{0.01(n-2)}=t_{0.01(10)}=3.169$。所以当 x=98 时，y 总体平均数的 95%和 99%置信区间分别为：

[2709.9772-2.228×17.6111, 2709.9772+2.228×17.6111]

[2709.9772-3.169×17.6111, 2709.9772+3.169×17.6111]

即[2670.7397,2749.2147]和[2654, 2765.7868]。

这说明雏鹅重为 98 克时，70 日龄总体平均重在[2670.7397, 2749.2147]区间内，其可靠度为 95%；在[2654, 2765.7868]区间内，其可靠度为 99%。

4、单个 y 值的置信区间　有时需要估计当 x 取某一数值时，相应 y 总体的一个 y 值的置信区间。因为 $(\hat{y}-y)/S_y$ 服从自由度为 $n-2$ 的 t 分布，其中，S_y 为单个 y 值的估计标准误，计算公式为：

$$S_y=S_{yx}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{SS_x}} \tag{8-21}$$

当 x 取某一数值时，单个 y 值的 95%、99%置信区间为：

$$[\hat{y}-t_{0.05(n-2)}S_y,\hat{y}+t_{0.05(n-2)}S_y] \tag{8-22}$$

$$[\hat{y}-t_{0.01(n-2)}S_y,\hat{y}+t_{0.01(n-2)}S_y] \tag{8-23}$$

【例 8.5】 求出**【例 8.1】**资料当 x=98 时单个 y 值的 95%和 99%置信区间。

对于**【例 8.1】**资料，当 x=98 时，$\hat{y}$=2709.9772, 且

$$S_y=S_{yx}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{SS_x}}=60.9525\times\sqrt{1+\frac{1}{12}+\frac{(98-98.5)^2}{1685}}=63.4457$$

$t_{0.05(n-2)}=t_{0.05(10)}=2.228$，$t_{0.01(n-2)}=t_{0.01(10)}=3.169$。所以当 x=98 时，某一 y 值的 95%和 99%置信区间分别为：

[2709.9772-2.228×63.4457, 2709.9772+2.228×63.4457]

[2709.9772-3.169×63.4457, 2709.9772+3.169×63.4457]

即[2568.6202,2851.3342]和[2508.9178, 2911.0366]。

这说明雏鹅重为 98 克时，就一只白鹅而言，70 日龄重在[2568.6202, 2851.3342]区间内，其可靠度为 95%；在[2508.9178, 2911.0366] 区间内，其可靠度为 99%。

从计算 $S_{\hat{y}}$ 的（8-18）式和 S_y 的（8-21）式可以看出：$S_{\hat{y}}$ 和 S_y 随 $(x-\bar{x})$ 的绝对值增大而增大，随 n 和 SS_x 的增大而减少。这表明，愈靠近 $\bar{x}$，对 y 总体平均值或单个 y 的估计值就愈精确，而增大样本含量，扩大 x 的取值范围亦可提高精确度。

第二节　直线相关

进行直线相关分析的基本任务在于根据 x、y 的实际观测值，计算表示两个相关变量 x、y 间线性相关程度和性质的统计量——相关系数 r 并进行显著性检验。

一、 决定系数和相关系数

在上一节中已经证明了等式：$\sum(y-\bar{y})^2 = \sum(\hat{y}-\bar{y})^2 + \sum(y-\hat{y})^2$ 。从这个等式不难看到：y 与 x 直线回归效果的好坏取决于回归平方和 $\sum(\hat{y}-\bar{y})^2$ 与离回归平方和 $\sum(y-\hat{y})^2$ 的大小，或者说取决于回归平方和在 y 的总平方和 $\sum(y-\bar{y})^2$ 中所占的比例的大小。这个比例越大，y 与 x 的直线回归效果就越好，反之则差。我们把比值 $\sum(\hat{y}-\bar{y})^2 / \sum(y-\bar{y})^2$ 叫做 x 对 y 的决定系数（**coefficient of determination**），记为 r^2，即

$$r^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} \qquad (8\text{-}24)$$

决定系数的大小表示了回归方程估测可靠程度的高低，或者说表示了回归直线拟合度的高低。显然有 $0 \leqslant r^2 \leqslant 1$。因为

$$r^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} = \frac{SP_{xy}{}^2}{SS_x SS_y} = \frac{SP_{xy}}{SS_x} \cdot \frac{SP_{xy}}{SS_y} = b_{yx} \cdot b_{xy}$$

而 SP_{xy}/SS_x 是以 x 为自变量、y 为依变量时的回归系数 b_{yx}。若把 y 作为自变量、x 作为依变量 ，则回归系数 $b_{xy}=SP_{xy}/SS_y$，所以决定系数 r^2 等于 y 对 x 的回归系数与 x 对 y 的回归系数的乘积。这就是说，决定系数反应了 x 为自变量、y 为依变量和 y 为自变量、x 为依变量时两个相关变量 x 与 y 直线相关的信息，即决定系数表示了两个互为因果关系的相关变量间直线相关的程度。但决定系数介于 0 和 1 之间，不能反应直线关系的性质——是同向增减或是异向增减。

若求 r^2 的平方根，且取平方根的符号与乘积和 SP_{xy} 的符号一致，即与 b_{xy} 、b_{yx} 的符号一致，这样求出的平方根既可表示 y 与 x 的直线相关的程度，也可表示直线相关的性质。统计学上把这样计算所得的统计量称为 x 与 y 的相关系数（**coefficient of correlation**），记为 r，即

$$r=\frac{SP_{xy}}{\sqrt{SS_xSS_y}} \tag{8-25}$$

$$=\frac{\sum xy-\frac{(\sum x)(\sum y)}{n}}{\sqrt{\left[\sum x^2-\frac{(\sum x)^2}{n}\right]\left[\sum y^2-\frac{(\sum y)^2}{n}\right]}} \tag{8-26}$$

二、相关系数的计算

【例 8.6】计算 10 只绵羊的胸围（*cm*）和体重(*kg*)（表 8-3）的相关系数。

表 8-3　10 只绵羊胸围和体重资料

编号	1	2	3	4	5	6	7	8	9	10
胸围(*x*)	68	70	70	71	71	71	73	74	76	76
体重(*y*)	50	60	68	65	69	72	71	73	75	77

根据表 8-3 所列数据先计算出：

$SS_x=\sum x^2-(\sum x)^2/n=51904-(720)^2/10=64$

$SS_y=\sum y^2-(\sum y)^2/n=46818-(680)^2/10=578$

$SP_{xy}=\sum xy-(\sum x)(\sum y)/n=49123-(720)(680)/10=163$

代入（8-25）式得：

$$r=\frac{SP_{xy}}{\sqrt{SS_x\cdot SS_y}}=\frac{163}{\sqrt{64\times578}}=0.8475$$

即绵羊胸围与体重的相关系数为 0.8475。

三、 相关系数的显著性检验

上述根据实际观测值计算得来的相关系数*r*是样本相关系数， 它是双变量正态总体中的总体相关系数 ρ 的估计值。样本相关系数*r*是否来自 ρ ≠0的总体，还须对样本相关系数*r*进行显著性检验。此时无效假设、备择假设为 H_0：ρ =0， H_A：ρ ≠0。 与直线回归关系显著性检验一样，可采用*t*检验法与F检验法对相关系数*r*的显著性进行检验。

t 检验的计算公式为

$$t=\frac{r}{S_r},\quad df=n-2 \tag{8-27}$$

其中， $S_r=\sqrt{(1-r^2)/(n-2)}$ ，叫做相关系数标准误。

F检验的计算公式为

$$F=\frac{r^2}{(1-r^2)/(n-2)},\quad df_1=1,\quad df_2=n-2 \tag{8-28}$$

统计学家已根据相关系数r显著性t检验法计算出了临界r值并列出了表格。 所以可以直接采用查表法对相关系数r进行显著性检验。具体作法是： 先根据自由度$n-2$查临界r值(附表8)，得$r_{0.05(n-2)}$，$r_{0.01(n-2)}$。 若$|r|<r_{0.05(n-2)}$，$P>0.05$，则相关系数r不显著，在r的右上方标记“ns”；若$r_{0.05(n-2)}\leqslant|r|<r_{0.01(n-2)}$，$0.01<P\leqslant0.05$，则相关系数$r$显著，在$r$的右上方标记“*”；若$|r|\geqslant r_{0.01(n-2)}$，$P\leqslant0.01$， 则相关系数$r$极显著，在$r$的右上方标记“**”。

对于【例8-6】，因为$df=n-2=10-2=8$，查附表8得：$r_{0.05(8)}=0.632$，$r_{0.01(8)}=0.765$，而$r=0.8475>r_{0.01(8)}$，$P<0.01$，表明绵羊胸围与体重的相关系数极显著。

四、相关系数与回归系数的关系

从相关系数计算公式的导出可以看到：相关变量 x 与 y 的相关系数 r 是 y 对 x 的回归系数与 x 对 y 的相关系数 b_{xy} 的几何平均数：

$$r=\sqrt{b_{yx}\cdot b_{xy}}$$

这表明直线相关分析与回归分析关系十分密切。事实上，它们的研究对象都是呈直线关系的相关变量。直线回归分析将二个相关变量区分为自变量和依变量，侧重于寻求它们之间的联系形式——直线回归方程；直线相关分析不区分自变量和依变量，侧重于揭示它们之间的联系程度和性质——计算出相关系数。两种分析所进行的显著性检验都是解决 y 与 x 间是否存在直线关系。因而二者的检验是等价的。即相关系数显著， 回归系数亦显著；相关系数不显著，回归系数也必然不显著。由于利用查表法对相关系数进行检验十分简便，因此在实际进行直线回归分析时，可用相关系数显著性检验代替直线回归关系显著性检验，即可先计算出相关系数 r 并对其进行显著性检验，若检验结果 r 不显著，则用不着建立直线回归方程；若 r 显著，再计算回归系数 b、回归截距 a，建立直线回归方程，此时所建立的直线回归方程代表的直线关系是真实的，可利用来进行预测和控制。

五、 应用直线回归与相关的注意事项

直线回归分析与相关分析在生物科学研究领域中已得到了广泛的应用，但在实际工作中却很容易被误用或作出错误的解释。为了正确地应用直线回归分析和相关分析这一工具，必须注意以下几点：

1、变量间是否存在相关 直线回归分析和相关分析毕竟是处理变量间关系的数学方法，在将这些方法应用于生物科学研究时要考虑到生物本身的客观实际情况，譬如变量间是否存在直线相关以及在什么条件下会发生直线相关，求出的直线回归方程是否有意义，某性状作为自变量或依变量的确定等等，都必须由生物科学相应的专业知识来决定，并且还要

用到生物科学实践中去检验。如果不以一定的生物科学依据为前提，把风马牛不相及的资料随意凑到一块作直线回归分析或相关分析，那将是根本性的错误。

2、其余变量尽量保持一致　由于自然界各种事物间的相互联系和相互制约，一个变量的变化通常会受到许多其它变量的影响，因此，在研究两个变量间关系时，要求其余变量应尽量保持在同一水平，否则，回归分析和相关分析可能会导致完全虚假的结果。例如人的身高和胸围之间的关系，如果体重固定，身高越高的人，胸围越小，但当体重在变化时，其结果就会相反。

3、观测值要尽可能的多　在进行直线回归与相关分析时，两个变量成对观测值应尽可能多一些，这样可提高分析的精确性，一般至少有 5 对以上的观测值。同时变量 x 的取值范围要尽可能大一些，这样才容易发现两个变量间的变化关系。

4、外推要谨慎　直线回归与相关分析一般是在一定取值区间内对两个变量间的关系进行描述，超出这个区间，变量间关系类型可能会发生改变，所以回归预测必须限制在自变量 x 的取值区间以内，外推要谨慎，否则会得出错误的结果。

5、正确理解回归或相关显著与否的含义　一个不显著的相关系数并不意味着变量 x 和 y 之间没有关系，而只有能说明两变量间没有显著的直线关系；一个显著的相关系数或回归系数亦并不意味着 x 和 y 的关系必定为直线，因为并不排除有能够更好地描述它们关系的非线性方程的存在。

6、一个显著的回归方程并不一定具有实践上的预测意义　如一个资料 x、y 两个变量间的相关系数 r=0.5，在 df=24 时，$r_{0.01(24)}$ =0.496，$r>r_{0.01(24)}$，表明相关系数极显著。而 r^2=0.25，即 x 变量或 y 变量的总变异能够通过 y 变量或 x 变量以直线回归的关系来估计的比重只占 25%，其余的 75%的变异无法借助直线回归来估计。

*第三节　曲线回归

一、　曲线回归分析概述

直线关系是两变量间最简单的一种关系。这种关系往往在变量一定的取值范围内成立，取值范围一扩大，散点图就明显偏离直线，此时两个变量间的关系不是直线而是曲线。例如，细菌的繁殖速率与温度关系，畜禽在生长发育过程中各种生理指标与年龄的关系，乳牛的泌乳量与泌乳天数的关系等都属这种类型。可用来表示双变量间关系的曲线种类很多，但许多曲线类型都可以通过变量转换化成直线形式，先利用直线回归的方法配合直线回归方程，然后再还原成曲线回归方程。

曲线回归分析（**curvilinear regression analysis**）的基本任务是通过两个相关变量 x 与 y 的实际观测数据建立曲线回归方程，以揭示 x 与 y 间的曲线联系的形式。

曲线回归分析最困难和首要的工作是确定变量与 x 间的曲线关系的类型。通常通过两个途径来确定：1、利用生物科学的有关专业知识，根据已知的理论规律和实践经验。例如，细菌数量的增长常具有指数函数的形式：$y = ae^{bx}$ ；幼畜体重的增长常具有“S”型曲线的形状，即 *Logistic* 曲线的形式等。2、若没有已知的理论规律和经验可资利用，则可用描点

法将实测点在直角坐标纸上描出，观察实测点的分布趋势与哪一类已知的函数曲线最接近，然后再选用该函数关系式来拟合实测点。

对于可直线化的曲线函数类型，曲线回归分析的基本过程是：先将 x 或 y 进行变量转换，然后对新变量进行直线回归分析——建立直线回归方程并进行显著性检验和区间估计，最后将新变量还原为原变量，由新变量的直线回归方程和置信区间得出原变量的曲线回归方程和置信区间。

还有一情况是找不到已知的函数曲线较接近实测点的分布趋势，这时可利用多项式回归，通过逐渐增加多项式的高次项来拟合，直到满意为止。该内容将在下一章的多项式回归中讨论。

二、能直线化的曲线类型

下面是几种常用的能直线化的曲线函数类型及其图型，并将其直线化，供进行曲线回归分析时选用。

1、双曲线函数 $1/y=a+b/x$

若令 $y'=1/y, x'=1/x$，则可将双曲线函数直线化为：$y'=a+bx'$

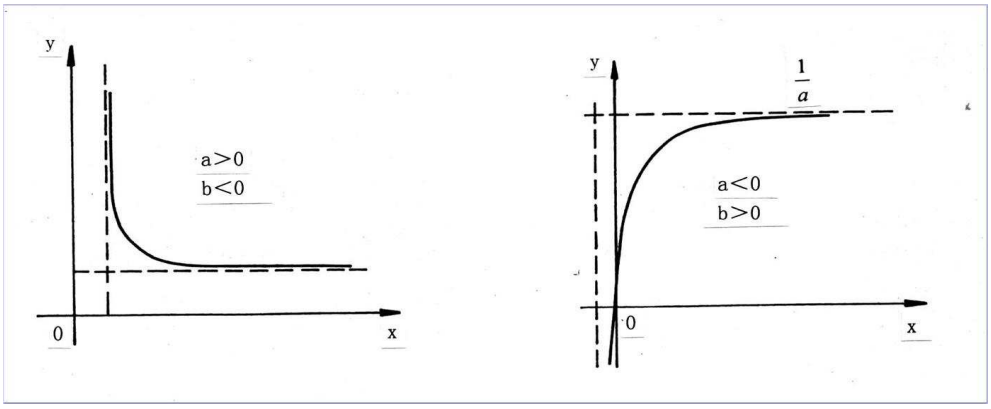


图 8-5 双曲线函数 $1/y=a+b/x$ 图形（虚线为渐进线）

2、幂函数 $y=ax^b$ (a>0)

若对幂函数 $y=ax^b$ 两端求自然对数，得：$\ln y=\ln a+\ln x$

并令 $y'=\ln y, a'=\ln a, x'=\ln x$，则可将幂函数直线化为：

$$y'=a'+bx'$$

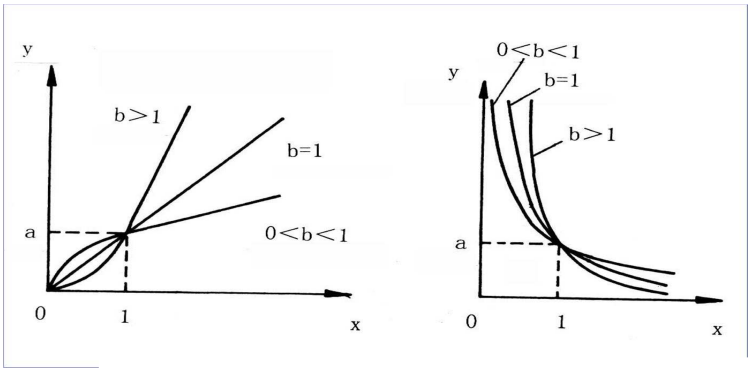


图 8-6 幂函数 $y=ax^b$（a>0）图形

3、指数函数 $y=ae^{bx}$ 或 $y=ae^{b/x}$（a>0）

（1）若对指数函数 $y=ax^{b}$（图 8-7a）两端求自然对数，得：$\ln y=\ln a+x$

并令 $y'=\ln y, a'=\ln a$，则可将其直线化为：$y'=a'+bx$

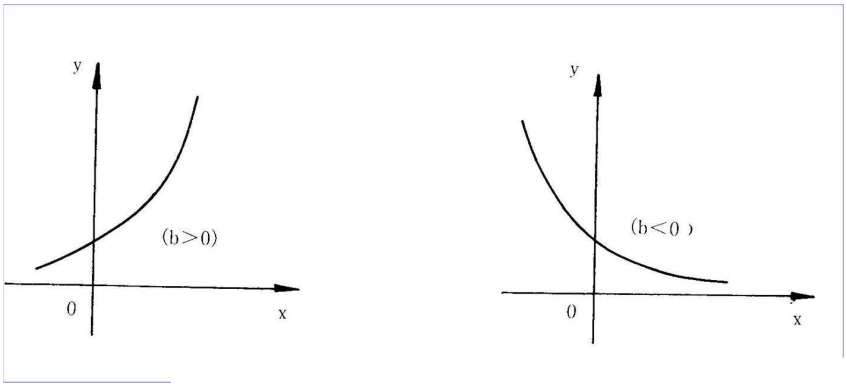
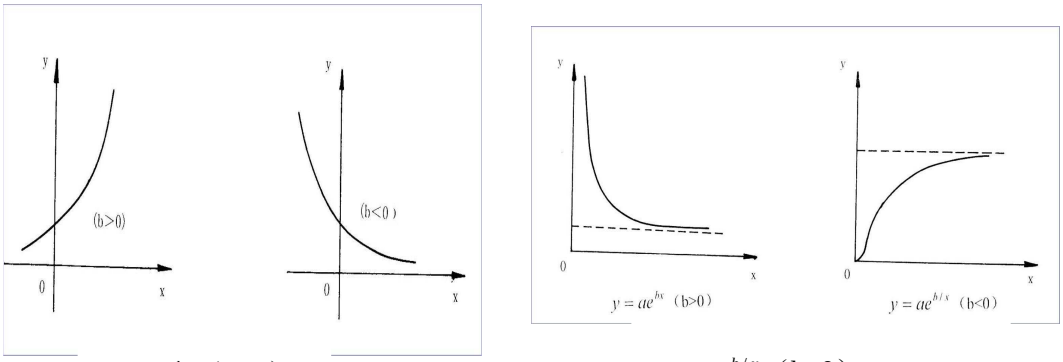


图 8- 7a　　指数函数 $y=ae^{bx}$ 图形

（2）若对指数函数 $y=ae^{b/x}$(图 8-7b)两端取自然对数，得：$\ln y=\ln a+b/x$

并令 $y'=\ln y, a'=\ln a, x'=1/x$，则可将其直线化为：$y'=a'+bx'$



$y=ae^{bx}$（b>0）　　$y=ae^{b/x}$（b<0）

图 8-7b　　指数函数 $y=ae^{b/x}$ 图形

4、对数函数 $y=a+b\lg x$

令 $x'=\lg x$，则将其直线化为 $y=a+bx'$

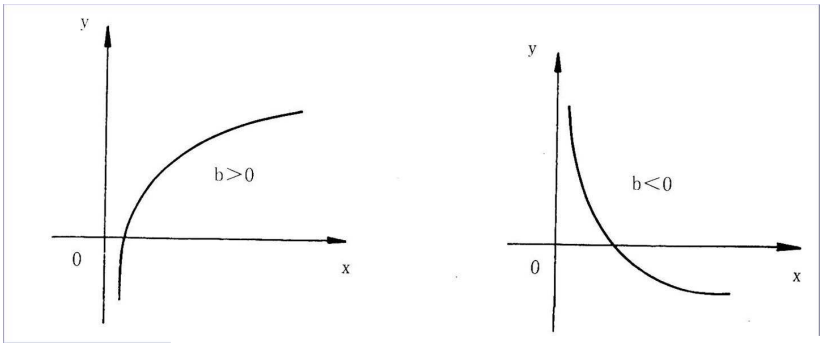


图 8-8 对数函数 $y=a+b\lg x$ 图

5、*Logistic* 生长曲线 $y=\dfrac{k}{1+ae^{-bx}}$

若将 *Logistic* 生长曲线两端取倒数，得：

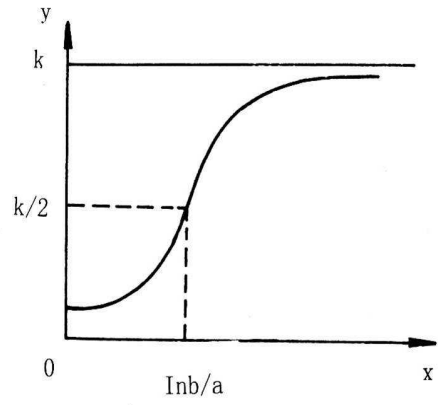


图 8-9 Logistic 生长曲线 $y=\dfrac{k}{1+ae^{-bx}}$ 图形

$$\frac{k}{y}=1+ae^{-bx}\text{，}\qquad\qquad\frac{k-y}{y}=ae^{-bx}$$

对两端取自然对数，得 $\ln\dfrac{k-y}{y}=\ln a-bx$

令 $y'=\ln\dfrac{k-y}{y},a'=\ln a,b'=-b$，可将其直线化为：

$$y'=a'+b'x$$

【例 8.7】 测定黑龙江雌性鲟鱼体长（*cm*）和体重（*kg*），结果如 8—4 表所示，试对鲟鱼体重与体长进行回归分析。

1、根据实际观测值在直角坐标纸上作散点图，选定曲线类型　此例的散点图见图 8-10。从散点图实测点的分布趋势看出它比较接近幂函数曲线图形，因而选用 $y=ax^b$ 来进行拟合。取 $x'=\lg x$，$y'=\lg y$，$a'=\lg a$ 则可将其直线化为：$y'=a'+bx'$。

2、对 x',y' 进行直线回归分析

表 8-4 鲟鱼体长与体重数据表

序号	体长（x）	体重（y）	$x'=\lg x$	$y'=\lg y$	$\hat{y}$	$y-\hat{y}$
1	70.70	1.00	1.8495	0	1.16305	-0.16306
2	98.25	4.85	1.9923	0.6857	3.86206	0.98794
3	112.57	6.59	2.0514	0.8189	6.34346	0.24654
4	122.48	9.01	2.0881	0.9547	8.62909	0.38091
5	138.46	12.34	2.1413	1.0913	13.49604	-1.15604
6	148.00	15.50	2.1703	1.1903	17.20854	-1.70854
7	152.00	21.25	2.1818	1.3274	18.96637	2.28363
8	162.00	22.11	2.2095	1.3446	23.92790	-1.81970

根据表 8-4 计算得：

$$\bar{x}' = \sum x' / n = 16.6841/8 = 2.087725 \qquad \bar{y}' = \sum y' / n = 7.4129/8 = 1.167350$$

$$SS_{x'} = \sum x'^2 - (\sum x')^2 / n = 34.8954 - 16.6841^2 / 8 = 2.35031281$$

$$SS_{y'} = \sum y'^2 - (\sum y')^2 / n = 8.2299 - 7.4129^2 / 8 = 1.16735804$$

$$SP_{x'y'} = \sum x'y' - (\sum x')(\sum y') / n = -1.64472315$$

x' 与 y' 的相关系数为：

$$r_{x'y'} = SP_{x'y'} / \sqrt{SS_{x'} SS_{y'}} = -0.9930^{**}$$

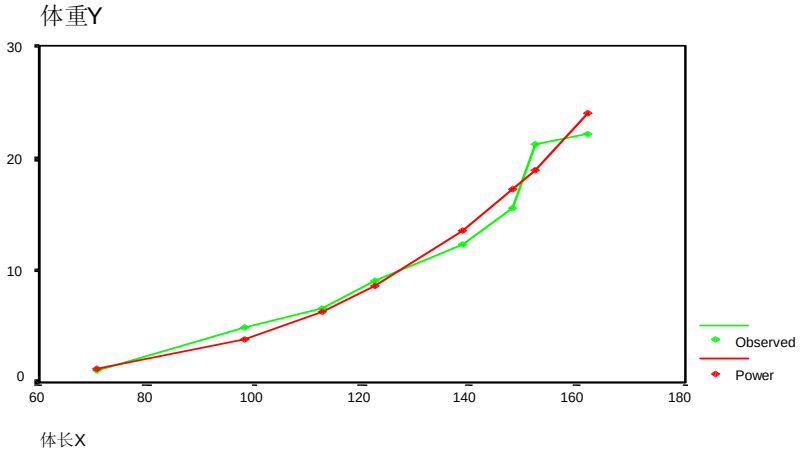


图 8-10 鲟鱼体长与体重散点图及回归曲线图

当 df=n-2=8-2=6 时，$r_{0.01(6)} = 0.834$，$|r| > r_{0.01(6)}$，P<0.01，表明 y' 与 x' 间存在极显著的线性关系。又因为：

$$b = SP_{x'y'} / SS_{x'} = 0.3669 / 0.1006 = 3.6417$$

$$a' = \bar{y}' - b\bar{x}' = 0.9266 - 3.6471 \times 2.0855 = -6.6794$$

得，y' 与 x' 的直线回归方程为：

$$\hat{y}' = -6.6797 + 3.6471x'$$

3、将变量 x'、y' 还原为 x、y

$$\lg \hat{y} = -6.6794 + 3.6471 \lg x$$

即：

$$\hat{y} = 2.0921 \times 10^{-7} x^{3.6471}$$

4、曲线配合的拟合度　曲线配合的好坏，即所配曲线与实测点吻合的好坏，取决于离回归平方和 $\Sigma(y-\hat{y})^2$ 与 y 的平方和 $\Sigma(y-\bar{y})^2$ 的比例大小。若这个比例小，说明所配曲线与实测点吻合程度高，反之则低，我们把数量 1 与这个比值之差定义为曲线回归的相关指数，记为 R^2，即：

$$R^2 = 1 - \frac{\Sigma(y-\hat{y})^2}{\Sigma(y-\bar{y})^2} \qquad (8\text{-}29)$$

相关指数 R^2 的大小表示了回归曲线拟合度的高低，或者说表示了曲线回归方程估测的可靠程度的高低。

对于【例 8. 7】先根据回归方程 $\hat{y}=2.0921\times10^{-7}x^{3.6471}$ 计算出各个回归估计值 $\hat{y}$ 和 $y-\hat{y}$，见表 8-4，计算出相关指数 R^2 为：

$$R^2=1-\frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}=1-\frac{13.98375}{409.0681}=0.9658$$

表明曲线回归方程 $\hat{y}=2.0921\times10^{-7}x^{3.6471}$ 的拟合度是比较高的，或者说该曲线回归方程估测的可靠程度比较高。

对于同一组实测数据，根据散点图的形状，可用几个相近的曲线进行拟合，同时建立几个曲线回归方程，此时可根据 R^2 的大小和生物学等专业知识，选择既符合生物学规律，拟合度又较高的曲线回归方程来描述这两个变量间的曲线关系。

【例 8. 8】在肉用四川白鹅的补饲料配方研究中，得到如下一组试验结果，试对体重与日龄进行回归分析。

表 8-5 肉用四川白鹅不同日龄的体重 （单位：*d*,*g*）

日龄(*x*)	体重(*y*)	(4316-*y*)/*y*	y'=*lg*[(4316-*y*)/*y*]	$\hat{y}$	*y*- $\hat{y}$
0	105	40.1048	1.6032	141.6726	-36.6726
7	214	19.1682	1.2826	212.2565	1.7435
14	335	11.8836	1.0749	315.3498	19.6502
21	560	6.7071	0.8265	462.8681	97.1319
28	790	4.4633	0.6479	667.8739	122.1261
42	1290	2.3457	0.3703	1287.6405	2.3595
56	2010	1.1473	0.0597	2144.4592	-134.4592
70	2950	0.4631	-0.3344	3005.5687	-55.5687

1、先根据实际观测数据在直角坐标纸上作散点图，选定曲线函数类型 此例的散点图见图 8-11。根据生物学知识和散点图的分布趋势，选用 *Logistic*(*S* 型)曲线，$y=\frac{k}{1+ae^{-bx}}$，其中 k 称为极限生长量。选取满足条件 $x_2=(x_1+x_3)/2$ 的 3 对观测值（14，335），（42，1290），（70，2950），算得 *k* 的估计值为：

$$k=\frac{y_2^2(y_1+y_3)-2y_1y_2y_3}{y_2^2-y_1y_3}$$

$$=\frac{1290^2(335+2950)-2\times335\times1290\times2950}{1290^2-335\times2950}=4316$$

因而在表 8-5 中得出(4316-*y*)/*y* 和 y'=*lg*[(4316-*y*)/*y*]

2、对 x 和 y' 进行直线回归分析

根据表 8-5，可算得：

$\bar{x}=\sum x/n=238/8=29.7500$　　$\bar{y}'=\sum y'/n=5.5325/8=06916$

$SS_x=\sum x^2-(\sum x)^2/n=11270-238^2/8=4189.5$

$SS_{y'}=\sum y'^2-(\sum y')^2/n=6.7284-5.5325^2/8=2.9024$

$SP_{xy'}=\sum xy'-(\sum x)(\sum y')/n=55.0627-238\times5.5325/8=-109.5292$

所以 x' 与 y' 的相关系数为：

$$r_{xy'} = SP_{xy'} / \sqrt{SS_x SS_{y'}} = -109.5292 / \sqrt{4189.5 \times 2.9024} = -0.9933^{**}$$

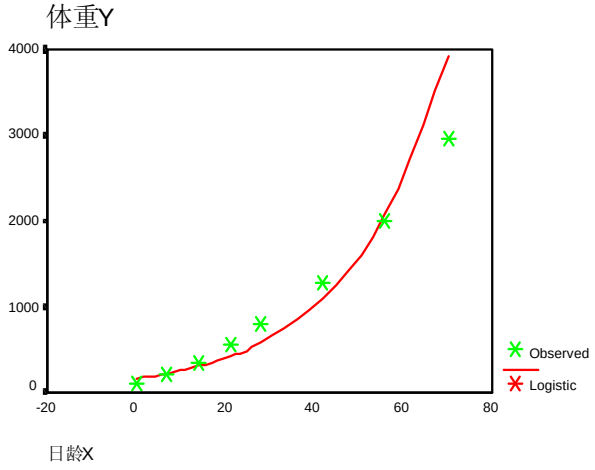


图 8-11 四川白鹅的日龄与体重散点图及回归曲线图

而当 *df*=*n*-2=8-2=6 时，$r_{0.01(6)} = 0.834$，$|r| > r_{0.01(6)}$，*P*<0.01，表明 *x* 与 *y* ´ 间存在极显著的直线关系。又因为：

$$b = SP_{xy'} / SS_x = -109.5292 / 4189.5 = -0.02614$$
$$a' = \bar{y}' - b\bar{x} = 0.6196 - (-0.02614) \times 29.75 = 1.4693$$

所以 y' 与 *x* 的直线回归方程为：

$$\hat{y}' = 1.4693 - 0.02614x$$

3、将变量 y′ 还原为 *y*

因为 $y' = \lg[(k-y)/y]$，$a' = \lg a$，$b' = -b\lg e$

所以 $a = 10^{a'} = 10^{1.4693} = 29.4646$

$$b = -b' / \lg e = -0.02614 / \lg e = 0.06019$$

因此 *Logistic*(*S* 型)生长曲线回归方程为：

$$\hat{y} = \frac{4316}{1 + 29.4646e^{-0.06019x}}$$

4、曲线配合的拟合度

先根据回归方程 $\hat{y} = \frac{4316}{1 + 29.4646e^{-0.06019x}}$ 计算出各个估计值 $\hat{y}$ 和 $y - \hat{y}$，见表 8-5，于是可以计算出相关指数 R^2 为：

$$R^2 = 1 - \frac{\Sigma(y-\hat{y})^2}{\Sigma(y-\bar{y})^2} = 1 - \frac{47256.164^2}{6997381.5} = 0.9932$$

表明曲线回归方程 $\hat{y} = \frac{4316}{1 + 29.4646e^{-0.06019x}}$ 的拟合度是比较高的，或者说该曲线回归方程估测的可靠程度是比较高的。

习　　题

1、什么叫直线回归分析？回归截距、回归系数与回归估计值 $\hat{y}$ 的统计意义是什么？

2、什么是直线相关分析？决定系数、相关系数的意义是什么？如何计算？

3、直线相关系数与回归系数的关系如何？直线相关系数与配合回归直线有何关系？

4、如何确定两个变量间的曲线类型？可直线化的曲线回归分析的基本步骤是什么？

5、10 头育肥猪的饲料消耗（x）和增重（y）资料如下表（单位：kg），试对增重与饲料消耗进行直线回归分析，并作出回归直线。

x	191	167	194	158	200	179	178	174	170	175
y	33	11	42	24	38	44	38	37	30	35

（r=0.6074，$\hat{y}$ =-47.8084+0.4536x）

6、试对下列资料进行直线相关和回归分析。

X	36	30	26	23	26	30	20	19	20	16
Y	0.89	0.80	0.74	0.80	0.85	0.68	0.73	0.68	0.80	0.58

（r=0.6369，$\hat{y}$ =0.5215+0.009492x）

7、对来航鸡胚胎生长的研究，测得 5—20 日龄鸡胚重量资料见下表，试建立鸡胚重依日龄变化的回归方程（用 *Logistic* 曲线拟合）

日龄 x（天）	5	6	7	8	9	10	11	12
胚重 y（g）	0.250	0.498	0.846	1.288	1.656	2.662	3.100	4.579

日龄 x（天）	13	14	15	16	17	18	19	20
胚重 y（g）	6.518	7.486	9.948	14.522	15.610	19.914	23.736	26.472

（$r_{xy'}$=0.9848，$\hat{y}$ =33.239/(1+659.9636e$^{-0.3863x}$)，R^2=0.9937）

第九章　多元线性回归与多项式回归

直线回归研究的是一个依变量与一个自变量之间的回归问题，但是，在畜禽、水产科学领域的许多实际问题中，影响依变量的自变量往往不止一个，而是多个，比如绵羊的产毛量这一变量同时受到绵羊体重、胸围、体长等多个变量的影响，因此需要进行一个依变量与多个自变量间的回归分析，即多元回归分析（**multiple regression analysis**），而其中最为简单、常用并且具有基础性质的是多元线性回归分析（**multiple linear regression analysis**），许多非线性回归（**non-linear regression**）和多项式回归（**polynomial regression**）都可以化为多元线性回归来解决，因而多元线性回归分析有着广泛的应用。研究多元线性回归分析的思想、方法和原理与直线回归分析基本相同，但是其中要涉及到一些新的概念以及进行更细致的分析，特别是在计算上要比直线回归分析复杂得多，当自变量较多时，需要应用电子计算机进行计算。

第一节　　多元线性回归分析

多元线性回归分析的基本任务包括：根据依变量与多个自变量的实际观测值建立依变量对多个自变量的多元线性回归方程；检验、分析各个自变量对依自变量的综合线性影响的显著性；检验、分析各个自变量对依变量的单纯线性影响的显著性，选择仅对依变量有显著线性影响的自变量，建立最优多元线性回归方程；评定各个自变量对依变量影响的相对重要性以及测定最优多元线性回归方程的偏离度等。

一、　多元线性回归方程的建立

（一）多元线性回归的数学模型　设依变量 y 与自变量 x_1 、 x_2 、…、 x_m 共有 n 组实际观测数据：

变量 / 序号	y	x_1	x_2	…	x_m
1	y_1	x_{11}	x_{21}	…	x_{m1}
2	y_2	x_{12}	x_{22}	…	x_{m2}
⋮	⋮	⋮	⋮	…	⋮
n	y_n	x_{1n}	x_{2n}	…	x_{mn}

假定依变量 y 与自变量 x_1、x_2、…、x_m 间存在线性关系，其数学模型为：

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + ... + \beta_m x_{mj} + \varepsilon_j \qquad (9\text{-}1)$$

（j=1,2,…,n）

式中，x_1、x_2、…、x_m 为可以观测的一般变量（或为可以观测的随机变量）；y 为可以观

测的随机变量，随 x_1、x_2、…、x_m 而变，受试验误差影响；ε_j 为相互独立且都服从 $N(0,\sigma^2)$ 的随机变量。我们可以根据实际观测值对 β_0、β_1、β_2、...、β_m 以及方差 σ^2 作出估计。

（二）建立线性回归方程 设 y 对 x_1、x_2、…、x_m 的 m 元线性回归方程为：

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_mx_m$$

其中的 b_0、b_1、b_2、…、b_m 为 β_0、β_1、β_2...、β_m 的最小二乘估计值。即 b_0、b_1、b_2、…、b_m 应使实际观测值 y 与回归估计值 $\hat{y}$ 的偏差平方和最小。

令 $Q = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2$

$$= \sum_{j=1}^{n}(y_j - b_0 - b_1x_{1j} - b_2x_{2j} - \cdots - b_mx_{mj})^2$$

Q 为关于 b_0、b_1、b_2、…、b_m 的 $m+1$ 元函数。

根据微分学中多元函数求极值的方法，若使 Q 达到最小，则应有：

$$\frac{\partial Q}{\partial b_0} = -2\sum_{j=1}^{n}(y_j - b_0 - b_1x_{1j} - b_2x_{2j} - \cdots - b_mx_{mj}) = 0$$

$$\frac{\partial Q}{\partial b_i} = -2\sum_{j=1}^{n}x_{ij}(y_j - b_0 - b_1x_{1j} - b_2x_{2j} - \cdots - b_mx_{mj}) = 0$$

（i =1、2、…、m）

经整理得：

$$\begin{cases} nb_0 + (\Sigma x_1)b_1 + (\Sigma x_2)b_2 + \cdots + (\Sigma x_m)b_m = \Sigma y \\ (\Sigma x_1)b_0 + (\Sigma x_1^2)b_1 + (\Sigma x_1x_2)b_2 + \cdots + (\Sigma x_1x_m)b_m = \Sigma x_1y \\ (\Sigma x_2)b_0 + (\Sigma x_2x_1)b_1 + (\Sigma x_2^2)b_2 + \cdots + (\Sigma x_2x_m)b_m = \Sigma x_2y \\ \vdots \quad\quad \vdots \quad\quad \vdots \quad \cdots \quad \vdots \quad\quad \vdots \\ (\Sigma x_m)b_0 + (\Sigma x_mx_1)b_1 + (\Sigma x_mx_2)b_2 + \cdots + (\Sigma x_m^2)b_m = \Sigma x_my \end{cases} \tag{9-2}$$

由方程组（9-2）中的第一个方程可得

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \cdots - b_m\bar{x}_m \tag{9-3}$$

即 $b_0 = \bar{y} - \sum_{i=1}^{m}b_i\bar{x}_i$

其中：$\bar{y} = \frac{1}{n}\sum_{j=1}^{n}y_j$，$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n}x_{ij}$

若记

$$SS_i = \sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2, \qquad SS_y = \sum_{j=1}^{n}(y_j - \bar{y})^2$$

$$SP_{ik} = \sum_{j=1}^{n}(x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k) = SP_{ki} \qquad SP_{io} = \sum_{j=1}^{n}(x_{ij} - \bar{x}_i)(y_j - \bar{y})$$

（i、k=1、2、…、m；$i \neq k$）

并将 $b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \cdots - b_m\bar{x}_m$ 分别代入方程组（9-2）中的后 m 个方程，经整理可得到关于偏回归系数 b_1、b_2、…、b_m 的正规方程组（**normal equations**）为：

$$\begin{cases} SS_1 b_1 + SP_{12} b_2 + \cdots + SP_{1m} b_m = SP_{10} \\ SP_{21} b_1 + SS_2 b_2 + \cdots + SP_{2m} b_m = SP_{20} \\ \vdots \quad\quad \vdots \quad \cdots \quad \vdots \quad\quad \vdots \\ \mathrm{SP_{m1}} b_1 + SP_{m2} b_2 + \cdots + SS_m b_m = SP_{m0} \end{cases} \tag{9-4}$$

解正规方程组（9-4）即可得偏回归系数b_1、b_2、…、b_m的解，而

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \cdots - b_m\bar{x}_m$$

于是得到m元线性回归方程

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m$$

m元线性回归方程的图形为$m+1$维空间的一个平面，称为回归平面；b_0称为回归常数项，当$x_1 = x_2 = \cdots = x_m = 0$时，$\hat{y} = 0$, 在$b_0$有实际意义时，$b_0$表示$y$的起始值；$b_i$（$i$=1、2、…、$m$）称为依变量$y$对自变量$x_i$的偏回归系数（**partial regression coefficient**），表示除自变量x_i以外的其余$m-1$个自变量都固定不变时，自变量x_i每变化一个单位，依变量y平均变化的单位数值，确切地说，当b_i>0 时，自变量x_i每增加一个单位，依变量y平均增加b_i个单位；当b_i<0时，自变量x_{i}每增加一个单位，依变量y平均减少b_i个单位。

若将$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \cdots - b_m\bar{x}_m$代入上式，则得

$$\hat{y} = \bar{y} + b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + \cdots + b_m(x_m - \bar{x}_m) \tag{9-5}$$

（9-5）式也为y对x_1、x_2、…、x_m的m元线性回归方程。

对于正规方程组（9-4），记

$$A = \begin{bmatrix} SS_1 & SP_{12} & \cdots & SP_{1m} \\ SP_{21} & SS_2 & \cdots & SP_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ SP_{m1} & SP_{m2} & \cdots & SS_m \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}, \quad B = \begin{bmatrix} SP_{10} \\ SP_{20} \\ \vdots \\ SP_{m0} \end{bmatrix}$$

则正规方程组（9-4）可用矩阵形式表示为

$$\begin{bmatrix} SS_1 & SP_{12} & \cdots & SP_{1m} \\ SP_{21} & SS_2 & \cdots & SP_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ SP_{m1} & SP_{m2} & \cdots & SS_m \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} SP_{10} \\ SP_{20} \\ \vdots \\ SP_{m0} \end{bmatrix} \tag{9-6}$$

即

$$Ab = B \tag{9-7}$$

其中A为正规方程组的系数矩阵、b为偏回归系数矩阵（列向量）、B为常数项矩阵（列向量）。

设系数矩阵A的逆矩阵为C矩阵，即$A^{-1} = C$，则

$$C = A^{-1} = \begin{bmatrix} SS_1 & SP_{12} & \cdots & SP_{1m} \\ SP_{21} & SS_2 & \cdots & SP_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ SP_{m1} & SP_{m2} & \cdots & SS_m \end{bmatrix}^{-1} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mm} \end{bmatrix}$$

其中：C矩阵的元素c_{ij}（i，j=1、2、…、m）称为高斯乘数，是多元线性回归分析中显著性检验所需要的。

关于求系数矩阵A的逆矩阵$\mathrm{A^{-1}}$的方法有多种，如行（或列）的初等变换法等，请参阅线性代数教材，这里就不再赘述。

对于矩阵方程（9-7）求解，有：

$$b = A^{-1}B$$
$$b = CB$$

即：

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} c_{11}\ c_{12} \cdots\ c_{1m} \\ c_{21}\ c_{22} \cdots\ c_{2m} \\ \vdots\ \ \vdots\ \ \cdots\ \vdots \\ c_{m1}\ c_{m2} \cdots\ c_{mm} \end{bmatrix} \begin{bmatrix} SP_{10} \\ SP_{20} \\ \vdots \\ SP_{m0} \end{bmatrix} \qquad (9\text{-}8)$$

关于偏回归系数b_1、b_2、…、b_m的解可表示为：

$$b_i = c_{i1}SP_{10} + c_{i2}SP_{20} + \cdots + c_{im}SP_{m0} \qquad (9\text{-}9)$$

（i=1、2、…、m）

或者 $b_i = \sum_{j=1}^{m} c_{ij}sp_{j0}$

而 $b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \cdots - b_m\bar{x}_m$

【例 9.1】 猪的瘦肉量是肉用型猪育种中的重要指标，而影响猪瘦肉量的有猪的眼肌面积、胴体长、膘厚等性状。设依变量y为瘦肉量（kg），自变量x_1为眼肌面积（cm^2），自变量x_2为胴体长（cm），自变量x_3为膘厚（cm）。根据三江猪育种组的54头杂种猪的实测数据资料，经过整理计算，得到如下数据：

$SS_1 = 846.2281$　$SS_2 = 745.6041$　$SS_3 = 13.8987$
$SP_{12} = 40.6832$　$SP_{13} = -6.2594$　$SP_{23} = -45.1511$
$SP_{10} = 114.4530$　$SP_{20} = 76.2799$　$SP_{30} = -11.2966$
$\bar{x}_1 = 25.7002$　$\bar{x}_2 = 94.4343$　$\bar{x}_3 = 3.4344$
$SS_y = 70.6617$　$\bar{y} = 14.8722$

试建立y对x_1、x_2、x_3的三元线性回归方程$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$。

将上述有关数据代入（9-5）式，得到关于偏回归系数b_1、b_2、b_3的正规方程组：

$$\begin{cases} 846.2281b_1 + 40.6832b_2 - 6.2594b_3 = 114.4530 \\ 40.6832b_1 + 745.6041b_2 - 45.1511b_3 = 76.2799 \\ -6.2594b_1 - 45.1511b_2 + 13.8987b_3 = -11.2966 \end{cases}$$

用线性代数有关方法求得系数矩阵的逆矩阵如下：

$$C = A^{-1}$$
$$= \begin{bmatrix} 846.2281 & 40.6832 & -6.2594 \\ 40.6832 & 745.6041 & -45.1511 \\ -6.2594 & -45.1511 & 13.8987 \end{bmatrix}^{-1}$$
$$= \begin{bmatrix} 0.001187 & -0.000040 & 0.000403 \\ -0.000040 & 0.001671 & 0.005410 \\ 0.000403 & 0.005410 & 0.089707 \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$$

根据式（9-8），关于b_1、b_2、b_3的解可表示为：

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} SP_{10} \\ SP_{20} \\ SP_{30} \end{bmatrix}$$

即关于b_1、b_2、b_3的解为：

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 0.001187 & -0.000040 & 0.000403 \\ -0.000040 & 0.001671 & 0.005410 \\ 0.000403 & 0.005410 & 0.089707 \end{bmatrix} \begin{bmatrix} 114.4530 \\ 76.2799 \\ -11.2966 \end{bmatrix} = \begin{bmatrix} 0.1282 \\ 0.0617 \\ -0.5545 \end{bmatrix}$$

而 $b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - b_3\bar{x}_3$

$$= 14.8722 - 0.1282 \times 25.7002 - 0.0617 \times 94.4343 - (-0.5545) \times 3.4344$$
$$= 7.6552$$

于是得到关于瘦肉量 y 与眼肌面积 x_1、胴体长 x_2、膘厚 x_3 的三元线性回归方程为：

$$\hat{y} = 7.6552 + 0.1282x_1 + 0.0617x_2 - 0.5545x_3$$

（三）多元线性回归方程的偏离度 以上根据最小二乘法，即使偏差平方和 $\sum(y-\hat{y})^2$ 最小建立了多元线性回归方程。偏差平方和 $\sum(y-\hat{y})^2$ 的大小表示了实测点与回归平面的偏离程度，因而偏差平方和又称为离回归平方和。统计学已证明，在 m 元线性回归分析中，离回归平方和的自由度为（n-m-1）。于是可求得离回归均方为 $\sum(y-\hat{y})^2$/（n-m-1）。离回归均方是模型（9-1）中 σ^2 的估计值。离回归均方的平方根叫离回归标准误，记为 $S_{y.12...m}$（或简记为 S_e），即

$$S_{y.12...m} = S_e = \sqrt{\sum(y-\hat{y})^2 \big/ (n-m-1)} \qquad (9\text{-}10)$$

离回归标准误 $S_{y.12...m}$ 的大小表示了回归平面与实测点的偏离程度，即回归估计值 $\hat{y}$ 与实测值 y 偏离的程度，于是我们把离回归标准误 $S_{y.12...m}$ 用来表示回归方程的偏离度。离回归标准误 $S_{y.12...m}$ 大，表示回归方程偏离度大，离回归标准误 $S_{y.12...m}$ 小，表示回归方程偏离度小。

利用公式 $\sum(y-\hat{y})^2$ 计算离回归平方和，因为先须计算出各个回归预测值 $\hat{y}$，计算量大，下面我们将介绍计算离回归平方和的简便公式。

二、多元线性回归的显著性检验

（一）多元线性回归关系的显著性检验 在畜禽、水产科学的许多实际问题中，我们事先并不能断定依变量 y 与自变量 x_1、x_2、…、x_m 之间是否确有线性关系，在根据依变量与多个自变量的实际观测数据建立多元线性回归方程之前，依变量与多个自变量间的线性关系只是一种假设，尽管这种假设常常不是没有根据的，但是在建立了多元线性回归方程之后，还必须对依变量与多个自变量间的线性关系的假设进行显著性检验，也就是进行多元线性回归关系的显著性检验，或者说对多元线性回归方程进行显著性检验。这里应用 F 检验方法。

与直线回归分析即一元线性回归分析一样，在多元线性回归分析中，依变量 y 的总平方和 SS_y 可以剖分为回归平方和 SS_R 与离回归平方和 SS_r 两部分，即：

$$SS_y = SS_R + SS_r \qquad (9\text{-}11)$$

依变量 y 的总自由度 df_y 也可以剖分为回归自由度 df_R 与离回归自由度 df_r 两部分，即：

$$df_y = df_R + df_r \qquad (9\text{-}12)$$

（9-11）与（9-12）两式称为多元线性回归的平方和与自由度的划分式或剖分式。

在（9-11）式中，$SS_y=\Sigma(y-\bar{y})^2$ 反映了依变量 y 的总变异；$SS_R=\Sigma(\hat{y}-\bar{y})^2$ 反映了依变量与多个自变量间存在线性关系所引起的变异，或者反映了多个自变量对依变量的综合线性影响所引起的变异；$SS_r=\Sigma(y-\hat{y})^2$ 反映了除依变量与多个自变量间存在线性关系以外的其他因素包括试验误差所引起的变异。

（9-11）式中各项平方和的计算方法如下：

$$\begin{aligned} SS_y &= \Sigma y^2-(\Sigma y)^2/n \\ SS_R &= b_1SP_{10}+b_2SP_{20}+\cdots+b_mSP_{m0}=\sum_{i=1}^{m}b_iSP_{io} \\ SS_r &= SS_y-SS_R \end{aligned} \qquad (9\text{-}12)$$

（9-12）式中各项自由度的计算方法如下：

$$df_y=n-1$$
$$df_R=m$$
$$df_r=n-m-1$$

在上述计算方法中，m 为自变量的个数，n 为实际观测数据的组数。

在计算出 SS_R、df_R 与 SS_r、df_r 之后，我们可以方便地算出回归均方 MS_R 与离回归均方 MS_r：

$$MS_R=\frac{SS_R}{df_R}；\quad MS_r=\frac{SS_r}{df_r}$$

检验多元线性回归关系是否显著或者多元线性回归方程是否显著，就是检验各自变量的总体偏回归系数 $\beta_i(i=1、2、\cdots、m)$ 是否同时为零，显著性检验的无效假设与备择假设为：

$H_0:\beta_1=\beta_2=\cdots=\beta_m=0, H_A:\beta_1、\beta_2、\cdots、\beta_m$不全为零

在 H_0 成立条件下，有

$$F=\frac{MS_R}{MS_r},\quad (df_1=df_R, df_2=df_r) \qquad (9\text{-}14)$$

由上述 F 统计量进行 F 检验即可推断多元线性回归关系的显著性。

这里特别要说明的是，上述显著性检验实质上是测定各自变量对依变量的综合线性影响的显著性，或者测定依变量与各自变量的综合线性关系的显著性。如果经过 F 检验，多元线性回归关系或者多元线性回归方程是显著的，则不一定每一个自变量与依变量的线性关系都是显著的，或者说每一个偏回归系数不一定都是显著的，这并不排斥其中存在着与依变量无线性关系的自变量的可能性。在上述多元线性回归关系显著性检验中，无法区别全部自变量中，哪些是对依变量的线性影响是显著的，哪些是不显著的。因此，当多元线性回归关系经显著检验为显著时，还必须逐一对各偏回归系数进行显著性检验，发现和剔除不显著的偏回归关系对应的自变量。另外，多元线性回归关系显著并不排斥有更合理的多元非线性回归方程的存在，这正如直线回归显著并不排斥有更合理的曲线回归方程存在一样。

对于【例 9.1】，建立的三元线性回归方程为：

$$\hat{y}=7.6552+0.1282x_1+0.0617x_2-0.5545x_3$$

现在对三元线性回归关系进行显著性检验。

已计算得：

$$SS_y=70.6617$$

而　　$SS_R = b_1SP_{10} + b_2SP_{20} + b_3SP_{30}$

$= 0.1282\times114.4530 + 0.0617\times76.2799 + (-0.5545)\times(-11.2966)$

$= 25.6433$

$SS_r = SS_y - SS_R$

$= 70.6617 - 25.6433$

$= 45.0184$

并且 df_y=n-1=54-1=53

df_R=m=3

df_r=n-m-1=54-3-1=50

列出方差分析表，进行 F 检验：

表 9-1　三元线性回归关系方差分析表

变异来源	SS	df	MS	F
回　归	25.6433	3	8.5478	9.493**
离回归	45.0184	50	0.9004	
总变异	70.6617	53		

由 df_1=3、df_2=50 查 F 值表得 $F_{0.01(3,50)}$=4.20, 因为 $F>F_{0.01(3,50)}$, $P<0.01$。表明，猪瘦肉量 y 与眼肌面积 x_1、胴体长 x_2、膘厚 x_3 之间存在极显著的线性关系，或者眼肌面积 x_1、胴体长 x_2、膘厚 x_3 对瘦肉量 y 的综合线性影响是极显著的。

（二）偏回归系数的显著性检验　当多元线性回归关系经显著性检验为显著或极显著时，还必须对每个偏回归系数进行显著性检验，以判断每个自变量对依变量的线性影响是显著的还是不显著的，以便从回归方程中剔除那些不显著的自变量，重新建立更为简单的多元线性回归方程。偏回归系数 b_i ($i=1$、2、…、m)的显著性检验或某一个自变量对依变量的线性影响的显著性检验所建立的无效假设与备择假设为：

$H_0:\beta_i = 0, H_A:\beta_i \neq 0$　　($i=1$、2、…、m)

有两种完全等价的显著性检验方法——t 检验与 F 检验。

1、t 检验

$$t_{b_i} = \frac{b_i}{S_{b_i}}, df = n-m-1, \quad (i=1、2、\cdots、m) \qquad (9\text{-}15)$$

式中 $S_{b_i} = S_{y\cdot12\cdots m}\cdot\sqrt{c_{ii}}$ 为偏回归系数标准误；

$S_{y\cdot12\cdots m} = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-m-1}} = \sqrt{MS_r}$ 为离回归标准误；

c_{ii} 为 C=A^{-1} 的主对角线元素。

2、F 检验　在多元线性回归分析中，回归平方和 SS_R 反映了所有自变量对依变量的综合线性影响，它总是随着自变量的个数增多而有所增加，但决不会减少。因此，如果在所考虑的所有自变量当中去掉一个自变量时，回归平方和 SS_R 只会减少，不会增加。减少的数值越大，说明该自变量在回归中所起的作用越大，也就是该自变量越重要。

设 SS_R 为 m 个自变量 x_1、x_2、…、x_m 所引起的回归平方和，SS'_R 为去掉一个自变量 x_i 后 m-1 个自变量所引起的回归平方和，那么它们的差 $SS_R - SS'_R$ 即为去掉自变量 x_i 之后，回归平方和所减少的量，称为自变量 x_i 的偏回归平方和，记为 SS_{b_i}，即：

$$SS_{b_i} = SS_R - SS'_R$$

可以证明：

$$SS_{b_i} = b_i^2 / c_{ii} \qquad (i=1、2、\cdots、m) \qquad (9\text{-}16)$$

偏回归平方和可以衡量每个自变量在回归中所起作用的大小，或者说反映了每个自变量对依变量的影响程度的大小。值得注意的是，在一般情况下，

$$SS_R \neq \sum_{i=1}^{m} SS_{b_i}$$

这是因为 m 个自变量之间往往存在着不同程度的相关，使得各自变量对依变量的作用相互影响。只有当 m 个自变量相互独立时，才有

$$SS_R = \sum_{i=1}^{m} SS_{b_i}$$

偏回归平方和 SS_{b_i} 是去掉一个自变量使回归平方和减少的部分，也可理解为添入一个自变量使回归平方和增加的部分，其自由度为 1，称为偏回归自由度，记为 df_{b_i}，即 $df_{b_i}=1$。显然，偏回归均方 MS_{b_i} 为

$$MS_{b_i} = SS_{b_i} / df_{b_i} = SS_{b_i} = b_i^2 / c_{ii} \qquad (i=1、2、\cdots、m) \qquad (9\text{-}17)$$

检验各偏回归系数显著性的 F 检验法应用下述 F 统计量：

$$F_{b_i} = MS_{b_i} / MS_r \,, (df_1 = 1, df_2 = n-m-1) \qquad (i=1、2、\cdots、m) \qquad (9\text{-}18)$$

可以将上述检验列成方差分析表的形式。

对于【例 9.1】，我们已经进行了三元线性回归关系的显著性检验，且结果为极显著的。现在对三个偏回归系数分别进行显著性检验。

t 检验法：

首先计算

$$S_{y\cdot 123} = \sqrt{MS_r} = \sqrt{0.9004} = 0.9489$$
$$S_{b_1} = S_{y\cdot 123}\sqrt{c_{11}} = 0.9489 \times \sqrt{0.001187} = 0.0327$$
$$S_{b_2} = S_{y\cdot 123}\sqrt{c_{22}} = 0.9489 \times \sqrt{0.001671} = 0.0388$$
$$S_{b_3} = S_{y\cdot 123}\sqrt{c_{33}} = 0.9489 \times \sqrt{0.089707} = 0.2842$$

然后计算各 *t* 统计量的值：

$$t_{b_1} = b_1 / S_{b_1} = 0.1282 / 0.0327 = 3.921$$
$$t_{b_2} = b_2 / S_{b_2} = 0.0617 / 0.0388 = 1.590$$
$$t_{b_3} = b_3 / S_{b_3} = -0.5545 / 0.2842 = -1.951$$

由 $df=n-m-1=50$ 查 t 值表得 $t_{0.05(50)} = 2.008, t_{0.01(50)} = 2.678$。因为 $|t_{b_1}| > t_{0.01(50)}$、$|t_{b_2}| < t_{0.05(50)}$、$|t_{b_3}| < t_{0.05(50)}$，所以偏回归系数 b_1 是极显著的，而偏回归系数 b_2、b_3 都是不显著的。

F 检验法：

首先计算各个偏回归平方和：

$$SS_{b_1} = b_1^2 / c_{11} = 0.1282^2 / 0.001187 = 13.8460$$
$$SS_{b_2} = b_2^2 / c_{22} = 0.0617^2 / 0.001671 = 2.2782$$
$$SS_{b_3} = b_3^2 / c_{33} = (-0.5545)^2 / 0.089707 = 3.4275$$

进而计算各个偏回归均方：

$$MS_{b_1} = SS_{b_1}/1 = 13.8460$$
$$MS_{b_2} = SS_{b_2}/1 = 2.2782$$
$$MS_{b_3} = SS_{b_3}/1 = 3.4275$$

最后计算各 F 的值：

$$F_{b_1} = MS_{b_1}/MS_r = 13.8460/0.9004 = 15.378^{**}$$
$$F_{b_2} = MS_{b_2}/MS_r = 2.2782/0.9004 = 2.530$$
$$F_{b_3} = MS_{b_3}/MS_r = 3.4275/0.9004 = 3.807$$

由 df_1=1，df_2=50 查 F 值表得 $F_{0.05(1,\ 50)}$=4.03, $F_{0.01(1,\ 50)}$=7.17。因为 $F_{b_1} > F_{0.01(1,\ 50)}$，$F_{b_2} < F_{0.05(1,\ 50)}$，$F_{b_3} < F_{0.05(1,\ 50)}$，因此偏回归系数 b_1 极显著，而偏回归系数 b_2、b_3 均不显著。这与 t 检验的结论是一致的。

也可以把上述偏回归系数显著性检验的 F 检验结果列成方差分析表的形式：

表 9-2　偏回归系数显著性检验方差分析表

变异来源	*SS*	*df*	*MS*	*F*
x_1 的偏回归	13.8460	1	13.8460	15.378**
x_2 的偏回归	2.2782	1	2.2782	2.530
x_3 的偏回归	3.4275	1	3.4275	3.807
离　回　归	45.0184	50	0.9004	

（三）自变量剔除与重新建立多元线性回归方程　当对显著的多元线性回归方程中各个偏回归系数进行显著性检验都为显著时，说明各个自变量对依变量的单纯影响都是显著的。若有一个或几个偏回归系数经显著性检验为不显著时，说明其对应的自变量对依变量的作用或影响不显著，或者说这些自变量在回归方程中是不重要的，此时应该从回归方程中剔除一个不显著的偏回归系数对应的自变量，重新建立多元线性回归方程，再对新的多元线性回归方程或多元线性回归关系以及各个新的偏回归系数进行显著性检验，直至多元线性回归方程显著，并且各个偏回归系数都显著为止。此时的多元线性回归方程即为最优多元线性回归方程（**the best multiple linear regression equation**）。

1、自变量的剔除　当经显著性检验有几个不显著的偏回归系数时，我们一次只能剔除一个不显著的偏回归系数对应的自变量，被剔除的自变量的偏回归系数，应该是所有不显著的偏回归系数中的 F 值（或 $|t|$ 值、或偏回归平方和）为最小者。这是因为自变量之间往往存在着相关性，当剔除某一个不显著的自变量之后，其对依变量的影响很大部分可以转加到另外不显著的自变量对依变量的影响上。如果同时剔除两个以上不显著的自变量，那就会比较多地减少回归平方和，从而影响利用回归方程进行估测的可靠程度。

2、重新进行少一个自变量的多元线性回归分析　我们一次剔除一个不显著的偏回归系数对应的自变量，不能简单地理解为只须把被剔除的自变量从多元线性回归方程中去掉就行了，这是因为自变量间往往存在相关性，剔除一个自变量，其余自变量的偏回归系数的数值将发生改变，回归方程的显著性检验、偏回归系数的显著性检验也都须重新进行，也就是说应该重新进行少一个自变量的多元线性回归分析。

设依变量 y 与自变量 x_1、x_2、…、x_m 的 m 元线性回归方程为：

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_mx_m$$

如果 x_i 为被剔除的自变量，则 m-1 元线性回归方程为：

$$\hat{y} = b_0' + b_1' + \cdots + b_{i-1}'x_{i-1} + b_{i+1}'x_{i+1} + \cdots + b_m'x_m \qquad (9\text{-}19)$$

我们可以应用前面介绍过的 m 元线性回归方程的建立方法根据实际观测数据建立 m-1 元线性回归方程，但是这需要重新进行大量的计算。下面介绍利用 m 元线性回归方程与 m-1 元线性回归方程的对应偏回归系 b_j 与 b'_j 的关系以及 m 元正规方程组系数矩阵逆矩阵 C 的元素与 m-1 元正规方程组系数矩阵逆矩阵 C' 的元素之间的关系建立 m-1 元线性回归方程的方法。

设关于 m-1 元线性回归方程（9-19）中的偏回归系 b'_1、b'_2、…、b'_{i-1}、b'_{i+1}、…、b'_m 的正规方程组系数矩陈的逆矩阵为 C'，其各元素为：

$$c'_{jk} \quad （j、k=1、2、\cdots、i\text{-}1、i\text{+}1、\cdots、m; j\neq i; k\neq i）$$

可以证明：

$$c'_{jk} = c_{jk} - \frac{c_{ji}c_{ki}}{c_{ii}} \tag{9-20}$$

式中 c_{jk}、c_{ji}、c_{ki}、c_{ii} 均为 m 元正规方程组系数矩阵逆矩阵 C 的元素。这样我们就非常方便地计算出新的 m-1 阶逆矩阵 C' 的各元素，以进行 m-1 元线性回归方程的偏回归系数 b'_j 的显著性检验。

还可以证明，m-1 元线性回归方程中的偏回归系数 b'_j 与 m 元线性回归方程中偏回归系数 b_j 之间有如下关系：

$$b'_j = b_j - \frac{c_{ij}}{c_{ii}}\cdot b_i \qquad （j=1、2、\cdots、i-1、i+1、\cdots、m） \tag{9-21}$$

（9-21）式说明了可以利用原来的 m 元线性回归方程中的偏回归系数和 m 元正规方程组系数矩阵的逆矩阵 C 的元素 c_{ij} 来计算剔除一个自变量之后新的 m-1 元线性回归方程中的各偏回归系数。

而新的 m-1 元线性回归方程中常数项 b'_0 由下式计算：

$$b'_0 = \bar{y} - b'_1\bar{x}_1 - \cdots - b'_{i-1}\bar{x}_{i-1} - b'_{i+1}\bar{x}_{i+1} - \cdots - b'_m\bar{x}_m \tag{9-22}$$

于是我们利用（9-21）和（9-22）式可以方便地算出新的 m-1 元线性回归方程中的各个偏回归系数及常数项，这样即建立了剔除一个自变量之后新的 m-1 元线性回归方程：

$$\hat{y} = b'_0 + b'_1x_1 + \cdots + b'_{i-1}x_{i-1} + b'_{i+1}x_{i+1} + \cdots + b'_mx_m$$

在重新建立 m-1 元线性回归方程之后，仍然需要对 m-1 元线性回归关系和偏回归系数 b'_j 进行显著性检验，方法同前，但一些统计量需要重新进行计算。对于 m-1 元线性回归方程（9-19）：

回归平方和 $SS_R = b'_1SP_{10} + \cdots + b'_{i-1}SP_{i-1,0} + b'_{i+1}SP_{i+1,0} + \cdots + b'_mSP_{m0}$

回归自由度 $df_R = m-1$

离回归平方和 $SS_r = SS_y - SS_R$

离回归自由度 $df_r = n-m$

对偏回归系数 b'_j 进行显著性检验时：

$t_{b'_j} = b'_j / S_{b'_j}, df = n-m$

$S_{b'_j} = S_{y\cdot 12\cdots i-1\ i+1\cdots m}\sqrt{c'_{jj}}$

$S_{y\cdot 12\cdots i-1\ i+1\cdots m} = \sqrt{\dfrac{MS_r}{n-m}}$，$\mathrm{MS_r}$ 为新的离回归均方。

而新的偏回归平方和为：$SS_{b'_j} = b'^{\,2}_j \big/ c'_{jj}$

$$F_{b_j'}=\frac{MS_{b_j'}}{MS_r}=\frac{SS_{b_j'}}{MS_r}=\frac{b_j'^{\,2}\big/c_{jj}'}{MS_r}\qquad (df_1=1,df_2=n-m)$$

上式中的 MS_r 仍为新的离回归均方。

重复上述步骤，直至回归方程显著以及各偏回归系数都显著为止，即建立了最优多元线性回归方程。

对于【例 9.1】，建立的三元线性回归方程为

$$\hat{y}=7.6552+0.1282x_1+0.0617x_2-0.5545x_3$$

经显著性检验，回归方程极显著，偏回归系数 b_1 极显著，而 b_2、b_3 都是不显著的。因为 $F_{b_2}<F_{b_3}$，所以剔除偏回归系数 b_2 对应的自变量 x_2（胴体长），重新建立瘦肉量 y 对眼肌面积 x_1、膘厚 x_3 的二元线性回归方程：

$$\hat{y}=b_0'+b_1'x_1+b_3'x_3$$

根据（9-21）式：

$$b_j'=b_j-\frac{c_{ij}}{c_{ii}}\cdot b_i$$

计算 b_1' 和 b_3'。这里 i=2，j=1，3。

$$\begin{aligned}b_1'&=b_1-\frac{c_{21}}{c_{22}}\cdot b_2\\&=0.1282-\frac{-0.000040}{0.001671}\times 0.0617=0.1297\end{aligned}$$

$$\begin{aligned}b_3'&=b_3-\frac{c_{23}}{c_{22}}\cdot b_2\\&=(-0.5545)-\frac{0.005410}{0.001671}\times 0.0617=-0.7544\end{aligned}$$

而 b_0' 由（9-22）式计算：

$$\begin{aligned}b_0'&=\bar{y}-b_1'\bar{x}_1-b_3'\bar{x}_3\\&=14.8722-0.1297\times 25.7002-(-0.7544)\times 3.4344\\&=14.1298\end{aligned}$$

于是重新建立的二元线性回归方程为：

$$\hat{y}=14.1298+0.1297x_1-0.7544x_3$$

现在对二元线性回归方程或者二元线性回归关系进行显著性检验。

已计算得：

$$\begin{aligned}SS_y&=70.6617\\SS_R&=b_1'SP_{10}+b_3'SP_{30}\\&=0.1297\times 114.4530+(-0.7544)\times(-11.2966)\\&=23.3667\\SS_r&=SS_y-SS_R\\&=70.6617-23.3667\\&=47.2950\\df_y&=n-1=53\\df_R&=2\\df_r&=df_y-df_R=51\end{aligned}$$

列出方差分析表，进行 F 检验：

表 9-3　二元线性回归关系方差分析表

变异来源	*SS*	*df*	*MS*	*F*
回　归	23.3667	2	11.6834	12.598**
离回归	47.2950	51	0.9274	
总变异	70.6617	53		

由 $df_1=2, df_2=51$ 应用线性内插法求临界 F 值,得 $F_{0.01(2,51)}=5.05$, 因为 $F>F_{0.01(2,51)}$, $P<0.01$, 表明二元线性回归关系或二元线性回归方程是极显著的。

下面对偏回归系数 b_1' 和 b_3' 进行显著性检验，这里应用 F 检验法：

首先应用（9-20）式

$$c'_{jk}=c_{jk}-\frac{c_{ji}c_{ki}}{c_{ii}}$$

计算关于 b_1'、b_3' 的正规方程组系数矩阵的逆矩阵 C' 的主对角线上的各元素，这里 i=2，j、k=1、3。

$$c'_{11}=c_{11}-\frac{c_{12}c_{12}}{c_{22}}=0.001187-\frac{(-0.000040)^2}{0.001671}=0.001186$$

$$c'_{33}=c_{33}-\frac{c_{32}c_{32}}{c_{22}}=0.089707-\frac{0.005410^2}{0.001671}=0.072192$$

下面计算偏回归平方和：

$$SS_{b_1'}=b_1'^2/c'_{11}=0.1297^2/0.001186=14.1839$$

$$SS_{b_3'}=b_3'^2/c'_{33}=(-0.7544)^2/0.072192=7.8834$$

列出方差分析表，进行 F 检验：

表 9-4　偏回归系数显著性检验方差分析表

变异来源	*SS*	*df*	*MS*	*F*
x_1 的偏回归	14.1839	1	14.1839	15.294**
x_3 的偏回归	7.8834	1	7.8834	8.500**
离　回　归	47.2950	51	0.9274	

由 $df_1=1, df_2=51$ 应用线性内插法求临界 F 值，得 $F_{0.01(1,51)}$=7.16, 因为 $F_{b_1'}$、$F_{b_3'}$ 均大于 $F_{0.01(1,51)}$,表明二元线性回归方程的偏回归系数 b_1' 和 b_3' 都是极显著的,或者说明眼肌面积 x_1、膘厚 x_3 分别对瘦肉量 y 的线性影响都是极显著的。

于是我们得到【例 9.1】的最优二元线性回归方程为：

$$\hat{y}=14.1298+0.1297x_1-0.7544x_3$$

回归方程表明：猪的瘦肉量与眼肌面积、膘厚有着极显著的线性回归关系。当膘厚性状保持不变时，眼肌面积性状每增加 $1cm^2$，瘦肉量平均增加 0.1297kg；而当眼肌面积性状保持不变时，膘厚性状每增加 1cm，瘦肉量平均减少 0.7544kg。

该回归方程的离回归标准误为：

$$S_{y\cdot 13}=\sqrt{MS_r}=\sqrt{0.9274}=0.9630$$

（四）自变量主次的判断　在实际应用中，我们经常需要对最优多元线性回归方程中的自变量进行主次判断，以便抓住主要矛盾，更好地解决实际问题。

1、标准偏回归系数（**standard partial regression coefficient**）的比较

定义

$$b_i^* = b_i \frac{S_i}{S_y} = b_i \sqrt{\frac{SS_i}{SS_y}}, (i = 1,2,...,m) \quad (9\text{-}23)$$

为第 i 个自变量 x_i 的标准偏回归系数。式中：S_i 为第 i 个自变量 x_i 的样本标准差，S_y 为依变量 y 的样本标准差。

标准偏回归系数为不带单位的相对数，其绝对值的大小可以衡量对应的自变量对依变量作用的相对重要性。标准偏回归系数又称“通径系数”，其应用请参阅本章第五节。

在多元线性回归分析中，在各自变量之间无显著相关的情况下，可以比较各标准偏回归系数绝对值的大小，大者，其对应的自变量对依变量的作用是主要的。

2、偏回归平方和的比较　在多元线性回归分析中，当自变量间存在着显著相关时，或者当无法判断各自变量间的相关性时，应比较各自变量的偏回归平方和 SS_{b_i}（i=1、2、…、m）的大小来判断各自变量对依变量影响的主次，凡是偏回归平方和大的自变量，其对依变量的作用一定是主要的。

对于【例 9.1】建立的最优二元线性回归方程：

$$\hat{y} = 14.1298 + 0.1297x_1 - 0.7544x_3$$

已算得 $SS_{b_1'} = 14.1839, \quad SS_{b_3'} = 7.8843$

因为 $SS_{b_1'} > SS_{b_3'}$，所以在上述二元线性回归方程中，自变量 x_1（眼肌面积）对依变量 y（瘦肉量）的影响是主要的。

*第二节　复相关分析

一、复相关的概念及意义

研究一个变量与多个变量的线性相关称为复相关分析（**analysis of multiple correlation**）。从相关分析角度来说，复相关中的变量没有依变量与自变量之分，但是在实际应用中，复相关分析经常与多元线性回归分析联系在一起，因此，复相关分析一般指依变量 y 与 m 个自变量 x_1、 x_2、…、x_m 的线性相关。

在多元线性回归分析中，如果 m 个自变量对依变量的回归平方和 SS_R 占依变量 y 的总平方和 SS_y 的比率越大，则表明依变量 y 和 m 个自变量的线性联系越密切，或者表明依变量 y 与 m 个自变量的线性相关越密切，因此定义：

$$R^2 = SS_R / SS_y \quad (9\text{-}24)$$

为 y 与 x_1、 x_2、…、x_m 的复相关指数，简称相关指数（**correlation index**）。

相关指数 R^2 表示多元线性回归方程的拟合度，或者说表示用多元线回归方程进行预测的可靠程度。显然，

$$0 \le R^2 \le 1$$

定义：

$$R = \sqrt{SS_R / SS_y} \quad (9\text{-}25)$$

为依变量 y 与 m 个自变量 x_1、 x_2、…、x_m 的复相关系数（**multiple correlation coefficient**）。

复相关系数表示 y 与 x_1、 x_2、…、x_m 的线性关系的密切程度，由于 $\hat{y}$ 包含了 x_1 、x_2、…、

x_m的综合线性影响，因此，y 与 x_1 、x_2、…、x_m的复相关系数也就相当于 y 与 $\hat{y}$ 的简单相关系数，即

$$R = r_{y\hat{y}} \tag{9-26}$$

复相关系数的取值范围为：$0 \le R \le 1$。在自由度一定时，R 愈近于 1，复相关愈密切；愈近于 0，愈不密切。

二、复相关系数的显著性检验

复相关系数的显著性检验也就是对 y 与 x_1、 x_2、…、x_m的线性关系的显著性检验，因此，复相关系数的显著性检验与相应的多元线性回归关系的显著性检验或多元线性回归方程的显著性检验是完全等价的。复相关系数 R 的显著性检验有两种方法——F 检验法与查表法。

（一）F 检验法 设 ρ 为 y 与 x_1 、x_2、…、x_m的总体复相关系数，F 检验的无效假设与备择假设为：

$H_0 : \rho = 0;$ $\quad H_A : \rho \neq 0$

由下述 F 统计量检验 R 的显著性：

$$F_R = \frac{R^2/m}{(l-R^2)/(n-m-1)}, (df_1 = m, df_2 = n-m-1) \tag{9-27}$$

注意：因为 $R^2 = \dfrac{SS_R}{SS_y}$，代入（9-27）式得

$$\begin{aligned} F_R &= \frac{SS_R / m}{SS_y(1-\dfrac{SS_R}{SS_y})/(n-m-1)} \\ &= \frac{SS_R / m}{SS_r /(n-m-1)} \\ &= \frac{MS_R}{MS_r} \end{aligned}$$

说明利用（9-27）式计算的 F_R 值实际上就是多元线性回归关系显著性检验——F 检验计算的 F 值，也就是说复相关系数的显著性检验与多元线性回归关系的显著性检验是完全等价的。

（二）查表法 对于（9－27）式，由于在 df_1、df_2 一定时，给定显著水平α的 F 值一定，因此，可计算出相应于显著水平α时的临界 R 值：

$$R = \sqrt{\frac{df_1 F}{df_1 F + df_2}}$$

并将其列成表。因此复相关系数显著性检验可用简便的查表法进行。

由 $df = n-m-1$ 和变量的总个数 M=m+1 查附表 8《r 和 R 的显著数值表》得临界 R 值：$R_{0.05(n-m-1,M)}$、 $R_{0.01(n-m-1,M)}$，将 R 与 $R_{0.05(n-m-1,M)}$、 $R_{0.01(n-m-1,M)}$ 比较：

若$R < R_{0.05(n-m-1,M)}, P > 0.05,$则$R$不显著;

若$R_{0.05(n-m-1,M)} \le R < R_{0.01(n-m-1,M)}, 0.01 < P \le 0.05,$则$R$显著;

若$R \ge R_{0.01(n-m-1,M)}, P \le 0.01,$则$R$极显著;

对于【例 9.1】,依变量 y（瘦肉量）与自变量 x_1（眼肌面积）、x_2（胴体长）、x_3（膘

厚）的复相关系数

$$R=\sqrt{SS_R/SS_y}=\sqrt{25.6433/70.6617}=0.6024$$

由于 $F_R=\dfrac{R^2/m}{(1-R^2)/(n-m-1)}=\dfrac{0.6024^2/3}{(1-0.6024^2)/(54-3-1)}=9.493^{**}$，（$F_{0.01(3,\ 50)}=4.20$）表明 R 极显著。注意，这里的 F_R 值与三元线性回归关系显著性检验的 F 值是相同的。

若用查表法，则由 $df=n-m-1=50$ 与 $M=m+1=3+1=4$ 查附表 8 得 $R_{0.01(50,\ 4)}=0.449$，因为 $R>R_{0.01(50,\ 4)}$，$P<0.01$,故 R 为极显著。

显著性检验结果表明，猪的瘦肉量与眼肌面积、胴体长、膘厚间存在极显著的复相关。

由于篇幅的限制，附表 8 仅列出了 $M=3$，4，5 的临界 R 值。若 $M>5$，则采用 F 检验或根据多元线性回归关系显著性检验的结果来推断复相关系数的显著性。

*第三节　偏相关分析

多个相关变量间的关系是较为复杂的，任何两个变量间常常存在不同程度的简单相关关系，但是这种相关关系又包含有其他变量的影响。因此简单相关分析即直线相关分析没有考虑其他变量对这两个变量的影响，简单相关分析实际上并不能真实反映两个相关变量间的相关关系。而只有消除了其他变量的影响之后，研究两个变量间的相关性，才能真实地反映这两个变量间相关的性质与密切程度。偏相关分析就是固定其他变量不变而研究某两个变量间相关性的统计分析方法。

一、偏相关系数的意义与计算

（一）偏相关系数的意义　在多个相关变量中，其他变量保持固定不变，所研究的两个变量间的线性相关称为偏相关（**partial correlation**）。

用来表示两个相关变量偏相关的性质与程度的统计量叫偏相关系数（**partial correlation coefficient**）。

根据被固定的变量个数可将偏相关系数分级，偏相关系数的级数等于被固定的变量的个数。

当研究 2 个相关变量 x_1、x_2 的关系时，用直线相关系数 r_{12} 表示 x_1 与 x_2 线性相关的性质与程度。此时固定的变量个数为 0，所以直线相关系数 r_{12} 又叫做零级偏相关系数。

当研究 3 个相关变量 x_1、x_2、x_3 的相关时，我们把 x_3 保持固定不变，x_1 与 x_2 的相关系数称为 x_1 与 x_2 的偏相关系数，记为 $r_{12\cdot3}$，类似地，还有偏相关系数 $r_{13\cdot2}$、 $r_{23\cdot1}$。这 3 个偏相关系数固定的变量个数为 1，所以都叫做一级偏相关系数。

当研究 4 个相关变量 x_1、x_2、x_3、x_4 的相关时，须将其中的 2 个变量固定不变，研究另外两个变量间的相关。即此时只有二级偏相关系数才真实地反映两个相关变量间线性相关的性质与程度。二级偏相关系数共有 $C_4^2=6$ 个：$r_{12\cdot34}$，$r_{13\cdot24}$，$r_{14\cdot23}$，$r_{23\cdot14}$，$r_{24\cdot13}$，$r_{34\cdot12}$。

一般，当研究 m 个相关变量 x_1、$x_{2、}$、…、x_m 的相关时，只有将其中的 m-2 个变量保持固定不变，研究另外两个变量的相关才能真实地反映这两个相关变量间的相关，即此时只有

m-2 级偏相关系数才真实地反映了这两个相关变量间线性相关的性质与程度。m-2 级偏相关系数共有 $C_m^2 = m(m-2)/2$ 个。x_i 与 x_j 的 m-2 级偏相关系数记为 $r_{ij.}(i,j=1,2,\ldots,m,i \neq j)$。

偏相关系数的取值范围为[-1，1]，即：$-1 \leqslant r_{ij.} \leqslant 1$。

（二）偏相关系数的计算

1、一级偏相关系数的计算　设三个相关变量 x_1、x_2、x_3 共有 n 组实测数据：

序号	x_1	x_2	x_3
1	x_{11}	x_{21}	x_{31}
2	x_{12}	x_{22}	x_{32}
⋮	⋮	⋮	⋮
n	x_{1n}	x_{2n}	x_{3n}

一级偏相关系数可由零级偏相关系数即直线相关系数计算，计算公式为：

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{32}^2)}} \qquad (9\text{-}28)$$

$$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{(1-r_{21}^2)(1-r_{31}^2)}}$$

2、二级偏相关系数的计算　设四个相关变量 x_1、x_2、x_3、x_4 共有 n 组实测数据：

序号	x_1	x_2	x_3	x_4
1	x_{11}	x_{21}	x_{31}	x_{41}
2	x_{12}	x_{22}	x_{32}	x_{42}
⋮	⋮	⋮	⋮	⋮
n	x_{1n}	x_{2n}	x_{3n}	x_{4n}

二级偏相关系数可由一级偏相关系数计算，计算公式为：

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1-r_{14.3}^2)(1-r_{24.3}^2)}}$$

$$r_{13.24} = \frac{r_{13.2} - r_{14.2}r_{34.2}}{\sqrt{(1-r_{14.2}^2)(1-r_{34.2}^2)}}$$

$$r_{14.23} = \frac{r_{14.2} - r_{13.2}r_{43.2}}{\sqrt{(1-r_{13.2}^2)(1-r_{43.2}^2)}}$$

$$r_{23.14} = \frac{r_{23.1} - r_{24.1}r_{34.1}}{\sqrt{(1-r_{24.1}^2)(1-r_{34.1}^2)}} \qquad (9\text{-}29)$$

$$r_{24.13} = \frac{r_{24.1} - r_{23.1}r_{43.1}}{\sqrt{(1-r_{23.1}^2)(1-r_{43.1}^2)}}$$

$$r_{34.12} = \frac{r_{34.1} - r_{32.1}r_{42.1}}{\sqrt{(1-r_{32.1}^2)(1-r_{42.1}^2)}}$$

3、m−2 级偏相关系数的计算　设 m 个相关变量 x_1、x_2、…、x_m 共有 n 组观测数据：

序号	x_1	x_2	…	x_m
1	x_{11}	x_{21}	…	x_{m1}
2	x_{12}	x_{22}	…	x_{m2}
⋮	⋮	⋮	⋮	⋮
n	x_{1n}	x_{2n}	…	x_{mn}

m-2 级偏相关系数的计算方法如下：

首先计算简单相关系数即直线相关系数 r_{ij}：

$$r_{ij}=\frac{SP_{ij}}{\sqrt{SS_iSS_j}}，（i,j=1、2、\cdots、m) \tag{9-30}$$

其中：$SP_{ij}=\sum(x_i-\bar{x}_i)(x_j-\bar{x}_j)$，$SS_i=\sum(x_i-\bar{x}_i)^2$，$SS_j=\sum(x_j-\bar{x}_j)^2$，并由简单相关系数 r_{ij} 组成相关系数矩阵 R：

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix} \tag{9-31}$$

然后求相关系数矩阵 R 的逆矩阵 C：

$$C=R^{-1}=\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mm} \end{bmatrix} \tag{9-32}$$

则相关变量 x_i 与 x_j 的 m-2 级偏相关系数 $r_{ij.}$ 的计算公式为：

$$r_{ij.}=\frac{-c_{ij}}{\sqrt{c_{ii}c_{jj}}} \qquad (i,j=1、2、\cdots、m;i\neq j) \tag{9-33}$$

二、偏相关系数的显著性检验

（一）t 检验法 设相关变量 x_i 与 x_j 的总体偏相关系数为 $\rho_{ij..}$，则对偏相关系数 $r_{ij.}$ 进行显著性检验的无效假设与备择假设为：

$H_o:\rho_{ij.}=0$，$H_A:\rho_{ij.}\neq 0$

t 检验公式为：

$$t_{r_{ij.}}=\frac{r_{ij.}}{S_{r_{ij.}}}=\frac{r_{ij.}}{\sqrt{(1-r_{ij.}^{\ 2})/(n-m)}},df=n-m \tag{9-34}$$

（9-34）式中，$S_{r_{ij.}}$ 为偏相关系数标准误，$S_{r_{ij.}}=\sqrt{\frac{1-r_{ij.}^{\ 2}}{n-m}}$；$n$ 为观测数据组数，m 为相关变量总个数。

注意，m 个相关变量的偏相关分析中的 m 指相关变量的总个数；m 元线性回归分析中的 m 指自变量的个数；这两种分析方法中的 m 所表达的意义是不同的。

（二）查表法 由 $df=n-m$ 及变量个数 2 查附表 8《r 和 R 显著数值表》得 $r_{0.05(n-m,2)}$，

$r_{0.01(n-m,2)}$。将偏相关系数的绝对值$|r_{ij.}|$与 $r_{0.05(n-m,2)}$、$r_{0.01(n-m,2)}$进行比较，即可作出统计推断。

【例 9.2】 对【例 9.1】资料进行偏相关分析。注意，此时相关变量总个数 m=4。

首先由【例 9.1】的 SS_1、SS_2、SS_3、SS_y、SP_{12}、SP_{13}、SP_{23}、SP_{10}、SP_{20}、SP_{30} 计算变量 y、x_1、x_2、x_3 间的简单相关系数：

$$r_{12}=\frac{SP_{12}}{\sqrt{SS_1SS_2}}=\frac{40.6832}{\sqrt{846.2281\times745.6041}}=0.0512$$

$$r_{13}=\frac{SP_{13}}{\sqrt{SS_1SS_3}}=\frac{-6.2594}{\sqrt{846.2281\times13.8987}}=-0.0577$$

$$r_{23}=\frac{SP_{23}}{\sqrt{SS_2SS_3}}=\frac{-45.1511}{\sqrt{745.6041\times13.8987}}=-0.4435$$

$$r_{10}=\frac{SP_{10}}{\sqrt{SS_1SS_y}}=\frac{114.4530}{\sqrt{846.2281\times70.6617}}=0.4680$$

$$r_{20}=\frac{SP_{20}}{\sqrt{SS_2SS_y}}=\frac{76.2799}{\sqrt{745.6041\times70.6617}}=0.3323$$

$$r_{30}=\frac{SP_{30}}{\sqrt{SS_3SS_y}}=\frac{-11.2966}{\sqrt{13.8987\times70.6617}}=-0.3605$$

相关系数矩阵 R 为：

$$R=\begin{bmatrix}r_{11}&r_{12}&r_{13}&r_{10}\\r_{21}&r_{22}&r_{23}&r_{20}\\r_{31}&r_{32}&r_{33}&r_{30}\\r_{01}&r_{02}&r_{03}&r_{00}\end{bmatrix}=\begin{bmatrix}1&0.0512&-0.0577&0.4680\\0.0512&1&-0.4435&0.3323\\-0.0577&-0.4435&1&-0.3605\\0.4680&0.3323&-0.3605&1\end{bmatrix}$$

然后求得相关系数矩阵 R 的逆矩阵 C 为：

$$C=R^{-1}=\begin{bmatrix}c_{11}&c_{12}&c_{13}&c_{10}\\c_{21}&c_{22}&c_{23}&c_{20}\\c_{31}&c_{32}&c_{33}&c_{30}\\c_{01}&c_{02}&c_{03}&c_{00}\end{bmatrix}=\begin{bmatrix}1.312941&0.107564&-0.127506&-0.696166\\0.107564&1.308967&0.473288&-0.314690\\-0.127506&0.473288&1.341733&0.386094\\-0.696166&-0.314690&0.386094&1.569564\end{bmatrix}$$

因为我们需要研究的是瘦肉量（y）与眼肌面积（x_1）、胴体长（x_2）、膘厚（x_3）的二级偏相关系数，由（9-33）式可以算得：

$$r_{01.23}=\frac{-c_{01}}{\sqrt{c_{00}c_{11}}}=\frac{-(-0.696166)}{\sqrt{1.569564\times1.312941}}=0.4850$$

$$r_{02.13}=\frac{-c_{02}}{\sqrt{c_{00}c_{22}}}=\frac{-(-0.314690)}{\sqrt{1.569564\times1.308967}}=0.2195$$

$$r_{03.12}=\frac{-c_{03}}{\sqrt{c_{00}c_{33}}}=\frac{-0.386094}{\sqrt{1.569564\times1.341733}}=-0.2661$$

现在对上述三个二级偏相关系数进行 t 检验：

$$t_{r_{01.23}}=\frac{r_{01.23}}{\sqrt{(1-r_{01.23}{}^2)/(n-m)}}=\frac{0.4850}{\sqrt{(1-0.4850^2)/(54-4)}}=3.922^{**}$$

$$t_{r_{02.13}}=\frac{r_{02.13}}{\sqrt{(1-r_{02.13}{}^2)/(n-m)}}=\frac{0.2195}{\sqrt{(1-0.2195^2)/(54-4)}}=1.591$$

$$t_{r_{03.12}}=\frac{r_{03.12}}{\sqrt{(1-r_{03.12}{}^2)/(n-m)}}=\frac{-0.2661}{\sqrt{[1-(-0.2661)^2]/(54-4)}}=-1.952$$

由 $df = n - m = 54 - 4 = 50$ 查 t 值表得 $t_{0.05(50)}$=2.008、$t_{0.01(50)}$=2.678,因为 $t_{r_{01.23}} > t_{0.01(50)}$，$p < 0.01$，所以 $r_{01.23}$ 为极显著；而 $t_{r_{02.13}} < t_{0.05(50)}$，$|t_{r_{03.12}}| < t_{0.05(50)}$，$p > 0.05$, 因此，$r_{02.13}$和$r_{03.12}$ 都是不显著的。

如用查表法对上述三个二级偏相关系数进行显著性检验，则由 $df = n - m = 54 - 4 = 50$ 以及变量个数为 2 查附表 8《r和R 显著数值表》得 $r_{0.05(50)} = 0.273$、$r_{0.01(50)} = 0.354$，因为$r_{01.23} > r_{0.01(50)}$ 而 $r_{02.13} < r_{0.05(50)}$、$|r_{03.12}| < r_{0.05(50)}$,所以$r_{01.23}$ 为极显著，$r_{02.13}$、$r_{03.12}$ 都是不显著的，这与 t 检验结论一致。

显著性检验结果表明，瘦肉量（y）与眼肌面积（x_1）呈极显著的正的偏相关，而瘦肉量（y）与胴体长（x_2）、膘厚（x_3）的偏相关均为不显著。

从以上分析中，我们看到简单相关系数 $r_{10} = 0.4680$、$r_{20} = 0.3323$、$r_{30} = -0.3605$，在数值上分别与相应的二级偏相关系 $r_{10.23}$、$r_{20.13}$、$r_{30.12}$ 是有差别的。经显著性检验，r_{10}、r_{30} 都是极显著的，r_{20} 是显著的，而这与对应的二级偏相关系数的显著性也是不完全一致的。造成偏相关系数与简单相关系数在数值上相差的原因就在于各自变量间的相关性。在多变量资料中，偏相关系数与简单相关系数在数值上可以相差很大，甚至有时连符号都可能相反。只有偏相关分析才能正确地表示两个变量间的线性相关的性质和程度，才真实反映了两变量间的本质联系。而简单相关分析则可能由于其他变量的影响，反映的两个变量间的关系只是非本质的表面联系，所以是不可靠的。因此，对多变量资料进行相关分析时，应进行偏相关分析。

*第四节　多项式回归

一、　多项式回归概念

研究一个依变量与一个或多个自变量间多项式的回归分析方法，称为多项式回归（**polynomial regression**）。如果自变量只有一个时，称为一元多项式回归；如果自变量有多个时，称为多元多项式回归。

一元 m 次多项式回归方程为：

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \cdots + b_m x^m \qquad (9\text{-}35)$$

二元二次多项式回归方程为：

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2 + b_5 x_1 x_2 \qquad (9\text{-}36)$$

在一元回归分析中，如果依变量 y 与自变量 x 的关系为非线性的，但是又找不到适当的函数曲线来拟合，则可以采用一元多项式回归。多项式回归的最大优点就是可以通过增加 x 的高次项对实测点进行逼近，直至满意为止。事实上，多项式回归可以处理相当一类非线性问题，它在回归分析中占有重要的地位，因为任一函数都可以分段用多项式来逼近。因此，在通常的实际问题中，不论依变量与其他自变量的关系如何，我们总可以用多项式回归来进行分析。

二、　多项式回归分析的一般方法

多项式回归问题可以通过变量转换化为多元线性回归问题来解决。

对于一元 m 次多项式回归方程（9-35），令 $x_1 = x$ 、 $x_2 = x^2$ 、…、 $x_m = x^m$ ，则（9-35）就转化为 m 元线性回归方程

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m$$

因此用本章第一节的方法就可解决多项式回归问题。需要指出的是，在多项式回归分析中，检验回归系数 b_i 是否显著，实质上就是判断自变量 x 的 i 次方项 x^i 对依变量 y 的影响是否显著。

对于二元二次多项式回归方程（9-36），令 $z_1 = x_1$、$z_2 = x_2$、$z_3 = x_1^2$、$z_4 = x_2^2$、$z_5 = x_1 x_2$，则（9-36）就转化为五元线性回归方程

$$\hat{y} = b_0 + b_1 z_1 + b_2 z_2 + b_3 z_3 + b_4 z_4 + b_5 z_5$$

但随着自变量个数的增加,多元多项式回归分析的计算量急剧增加。多元多项式回归属于多元非线性回归问题，在这里不作介绍。

在多项式回归中较为常用的是一元二次多项式回归和一元三次多项式回归，下面结合一实例对一元二次多项式回归作详细介绍。

三、 一元二次多项式回归分析

【例 9.3】 给动物口服某种药物 A 1000mg，每间隔 1 小时测定血药浓度（g/ml），得到表 9-5 的数据（血药浓度为 5 头供试动物的平均值）。试建立血药浓度（依变量 y）对服药时间（自变量 x）的回归方程。

表 9-5 血药浓度与服药时间测定结果表

服药时间 x（小时）	1	2	3	4	5	6	7	8	9
血药浓度 y(g/ml)	21.89	47.13	61.86	70.78	72.81	66.36	50.34	25.31	3.17
$\hat{y}$	22.7182	46.2563	62.2684	70.7545	71.7146	65.1487	51.0568	29.4389	0.2950
y- $\hat{y}$	-0.8282	0.8737	-0.4084	0.0255	1.0954	1.2113	-0.7168	-4.1298	2.8750

（一）根据表 9-5 的数据资料绘制 x 与 y 的散点图 (见图 9-1)。由散点图我们看到：血药浓度最大值出现在服药后 5 小时，在 5 小时之前血药浓度随时间的增加而增加，在 5 小时之后随着时间的增加而减少，散点图呈抛物线形状，因此我们可以选用一元二次多项式来描述血药浓度与服药时间的关系，即进行一元二次多项式回归或抛物线回归。

图 9-1 表 9-5 资料的散点图

（二）进行变量转换 设一元二次多项式回归方程为：

$$\hat{y} = b_0 + b_1 x + b_2 x^2$$

令 $x_1 = x$ 、 $x_2 = x^2$ ，则得二元线性回归方程

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

（三）进行二元线性回归分析 先计算得：

$$\sum x_1 = \sum x = 45, \quad \sum x_2 = \sum x^2 = 285, \sum y = 419.65$$

$$\sum x_1^2 = \sum x^2 = 285, \quad \sum x_2^2 = \sum x^4 = 15333, \sum y^2 = 24426.5833$$

$$\sum x_1x_2=\sum x^3=2025,\sum x_1y=\sum xy=1930.45,\sum x_2y=\sum x^2y=10452.11$$

再计算得：

$$\mathrm{SS}_1=60.0000,\quad \mathrm{SS}_2=6308.0000,\quad \mathrm{SS}_y=4859.2364$$
$$\mathrm{SP}_{12}=600.0000,\quad \mathrm{SP}_{10}=-167.8000,\quad \mathrm{SP}_{20}=-2836.8067$$
$$\bar{x}_1=5.0000,\quad \bar{x}_2=31.6667,\quad \bar{y}=46.6278$$

于是得到关于b_1、b_2的正规方程组为：

$$\begin{cases}60.0000b_1+600.0000b_2=-167.8000\\600.0000b_1+6308.0000b_2=-2836.8067\end{cases}$$

求出上述正规方程组系数矩阵的逆矩阵为：

$$C=\begin{bmatrix}c_{11}&c_{12}\\c_{21}&c_{22}\end{bmatrix}=\begin{bmatrix}0.341349&-0.032468\\-0.032468&0.003247\end{bmatrix}$$

关于b_1、b_2的解为：

$$\begin{bmatrix}b_1\\b_2\end{bmatrix}=\begin{bmatrix}c_{11}&c_{12}\\c_{21}&c_{22}\end{bmatrix}\begin{bmatrix}SP_{10}\\SP_{20}\end{bmatrix}$$
$$=\begin{bmatrix}0.341349&-0.032468\\-0.032468&0.003247\end{bmatrix}\begin{bmatrix}-167.8000\\-2836.8067\end{bmatrix}$$
$$=\begin{bmatrix}34.8271\\-3.7630\end{bmatrix}$$

即：$b_1=34.8217,\ \mathrm{b}_2=-3.7630$

而$b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2=46.6278-34.8271\times5-(-3.7630)\times31.6667=-8.3459$

于是得到二元线性回归方程为：

$$\hat{y}=-8.3459+34.8271x_1-3.7630x_2$$

现在对二元线性回归方程或二元线性回归关系进行显著性检验。

$$SS_y=4859.2364$$
$$SS_R=b_1SP_{10}+b_2SP_{20}$$
$$SS_r=SS_y-SS_R=4859.2364-4830.9162=28.3202$$
$$df_y=n-1=9-1=8,df_R=2,df_r=df_y-df_R=8-2=6$$

列出方差分析表，进行 F 检验。

表 9-6　二元线性回归关系方差分析表

变异来源	*SS*	*df*	*MS*	*F*
回　归	4830.9162	2	2415.4581	511.750**
离回归	28.3202	6	4.7200	
总变异	4859.2364	8		

由$df_1=2,df_2=6$查 F 值表得$F_{0.01(2,6)}=10.92$，因为 $F>F_{0.01(2,6)}$，$P<0.01$,表明二元线性回归关系是极显著的。

偏回归系数b_1、b_2的显著检验，应用 F 检验法：

$$SS_{b_1}=b_1^{\,2}/c_{11}=34.8271^2/0.341349=3553.3337$$
$$SS_{b_2}=b_2^{\,2}/c_{22}=(-3.7630)^2/0.003247=4361.0006$$
$$F_{b_1}=\frac{MS_{b_1}}{MS_r}=\frac{SS_{b_1}/1}{MS_r}=\frac{3553.3337}{4.7200}=752.825**$$

$$F_{b_2}=\frac{MS_{b_2}}{MS_r}=\frac{SS_{b_2}/1}{MS_r}=\frac{4361.0006}{4.7200}=923.941^{**}$$

由 $df_1=1, df_2=6$ 查 F 值表得 $F_{0.01(1,6)}=13.47$，因为 $F_{b_1}>F_{0.01(1,6)}$、$F_{b_2}>F_{0.01(1,6)}$，表明偏回归系数 b_1 和 b_2 都是极显著的。

（四）建立一元二次多项式回归方程 将 x_1 还原为 x，x_2 还原为 x^2，即得 y 对 x 的一元二次多项式回归方程为：

$$\hat{y}=-8.3459+34.8271x-3.7630x^2$$

（五）计算相关指数 R^2 因为 $\sum(y-\hat{y})^2=33.1111$，$\sum(y-\bar{y})^2=4859.2364$，相关指数 R^2 为：

$$R^2=1-\frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}=0.9932$$

表明 y 对 x 的一元二次多项式回归方程的拟合度是比较高的，或者说该回归方程估测的可靠程度是比较高的。

*第五节　通径分析

在研究多个相关变量间的线性关系时，除了可以采用多元线性回归分析和偏相关分析，还可以采用通径分析（**path analysis**）。由 S・Wright(1921)提出，并经遗传育种工作者不断完善和改进的通径分析，在研究多个相关变量间关系中具有精确、直观等优点，在遗传育种工作中广泛应用于研究遗传相关、近交系数、亲缘系数、遗传力，确定综合选择指数、复合育种值，剖分性状间的相关系数为直接作用与间接作用的代数和等等。

一、通径系数（path coefficient）与决定系数

（一）通径、相关线与通径图 为直观起见，先讨论一个依变量、两个自变量的情况。

设三个相关变量 y 与 x_1、x_2 间存在线性关系，y 为依变量（结果），x_1、x_2 为自变量（原因）且彼此相关，回归方程为：

$$\hat{y}=b_0+b_1x_1+b_2x_2 \qquad (9\text{-}37)$$

或

$$y=b_0+b_1x_1+b_2x_2+e \qquad (9\text{-}38)$$

其中 e 为剩余项。

可用图 9-2 来表示三个相关变量间的关系。

在图 9-2 中，单箭头线“←——”表示变量间存在着因果关系，方向为由原因到结果，称为通径（**path**），也称为直接通径。双箭头线“↙↘”表示变量间存在着平行关系（互为因果），称为相关线（**correlation line**），一条相关线相当于两条尾端相联的通

径。将包含两条或两条以上通径、也可以包含一条相关线的链称为间接通径。如图 9-2 中，$x_1 \rightarrow y$ 为通径

图 9-2　自变量 x_1、x_2 与依变量 y 的通径图

或直接通径，$x_1 \longleftrightarrow x_2 \longrightarrow y$ 为间接通径。这种用来表示相关变量间因果关系与平行关系的箭形图称为通径图（**path chart**）。

（二）通径系数　通径图直观、形象地表达了相关变量间的关系，仅定性地表达还不够，还须进一步用数量表示因果关系中原因对结果影响的相对重要程度与性质、平行关系中变量间相关的相对重要程度与性质，也就是必须用数量表示“通径”与“相关线”的相对重要程度与性质。

表示“通径”相对重要程度与性质的数量叫通径系数。表示“相关线”相对重要程度与性质的数量叫相关系数。

相关系数已在第八章进行了详细介绍。下面介绍通径系数的数学表达式。

设依变量 y 与自变量 x_1、x_2 间存在线性关系，回归方程为：

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

或

$$y = b_0 + b_1 x_1 + b_2 x_2 + e$$

其中 e 为剩余项：$e = y - \hat{y}$，且 $\sum e = 0, \bar{e} = 0$；x_1、x_2 彼此相关。表示这三个相关变量间关系的通径图见图 9-2。

由于偏回归系数 b_1、b_2 是带有单位的，一般不能直接由 b_1、b_2 比较自变量 x_1、x_2（原因）对依变量 y（结果）影响的重要程度的大小。为了能直接比较各自变量对依变量影响重要程度的大小，现将 y、x_1、x_2 三个变量及剩余项 e 进行标准化变换，使 y、x_1、x_2 及 e 变为不带单位的相对数。

由 $y = b_0 + b_1 x_1 + b_2 x_2 + \mathrm{e}$　可得

$$\bar{y} = b_0 + b_1 \bar{x}_1 + b_2 \bar{x}_2 \qquad (9\text{-}39)$$

将上述两式等号左右两端相减得

$$y - \bar{y} = b_1 (x_1 - \bar{x}_1) + b_2 (x - \bar{x}_2) + e \qquad (9\text{-}40)$$

再将（9-40）式两端同除以 y 的标准差 S_0，并作相应的恒等变形，得：

$$\frac{y - \bar{y}}{S_0} = b_1 \frac{s_1}{s_0} \cdot \frac{x_1 - \bar{x}_1}{s_1} + b_2 \frac{s_2}{s_0} \cdot \frac{x_2 - \bar{x}_2}{s_2} + \frac{s_e}{s_0} \cdot \frac{e}{s_e} \qquad (9\text{-}41)$$

式中：s_1、s_2、s_e 分别为 x_1、x_2 与 e 的标准差。

$b_1 \frac{s_1}{s_0}$、$b_2 \frac{s_2}{s_0}$ 为变量 x_1、x_2 标准化之后的偏回归系数，分别表示 x_1、x_2 对 y 影响的相对重要程度和性质；$\frac{S_e}{S_0}$ 表示剩余项 e 对 y 影响的相对重要程度和性质；$b_1 \frac{s_1}{s_0}$、$b_2 \frac{s_2}{s_0}$ 和 $\frac{S_e}{S_0}$ 就是 x_1、x_2 和 e 到 y 的通径系数的数学表达式。若把 x_1、x_2 和 e 到 y 的通径系数记为 $P_{0\cdot1}$、$P_{0\cdot2}$、$P_{0\cdot e}$，则有：

$$p_{0.1} = b_1 \frac{s_1}{s_0}, \qquad p_{0.2} = b_2 \frac{s_2}{s_0}, \qquad p_{0.e} = \frac{s_e}{s_0}$$

一般，若依变量 y 与自变量 x_1、x_2、…、x_m 间存在线性关系，回归方程为：

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m \mathrm{x_m} \qquad (9\text{-}42)$$

或 $$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m \mathrm{x_m} + \mathrm{e} \qquad (9\text{-}43)$$

当自变量两两相关时，其通径图如图 9-3 所示。则原因 $x_i\,(i = 1、2、\cdots、m)$ 与剩余项 e 到结果 y 的通径系数为：

$$p_{0.i} = b_i \frac{s_i}{s_0}, \quad p_{0.e} = \frac{s_e}{s_0}$$

图 9-3　自变量 x_1、x_2、…、x_m 与依变量 y 的通径图

（三）决定系数　通径系数的平方称为决定系数（**determination coefficient**）。决定系数表示原因（自变量）或误差对结果（依变量）的相对决定程度。

对于 $y = b_0 + b_1 x_1 + b_2 x_2 + \mathrm{e}$ 情况，原因 x_1、x_2 和剩余项 e 对结果 y 的决定系数分别记为 $d_{0.1}$、$d_{0.2}$、$d_{0.e}$，则：

$$d_{0.1} = {p_{0.1}}^2 = (b_1 \frac{s_1}{s_0})^2$$

$$d_{0.2} = {p_{0.2}}^2 = (b_2 \frac{s_2}{s_0})^2$$

$$d_{0.e} = {p_{0.e}}^2 = (\frac{s_e}{s_0})^2$$

对于 $y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m \mathrm{x_m} + \mathrm{e}$，原因 x_i（$i = 1、2、\cdots、m$）和剩余项 e 对结果 y 的决定系数记为 $d_{0.i}(i=1,2,\ldots,m)$和 $d_{0.\mathrm{e}}$。

则 $$d_{0.i} = {p_{0.i}}^2 = (b_i \frac{s_i}{s_0})^2 \quad , \quad d_{0.e} = {p_{0.e}}^2 = (\frac{s_e}{s_0})^2$$

二、通径系数的性质

可以证明通径系数有如下四个重要性质。

性质 1　如果相关变量 y、x_1、x_2 间存在线性关系，其中 y 为依变量（结果）、x_1 和 x_2 为自变量（原因），且 x_1 和 x_2 彼此相关，回归方程为

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

或 $$y = b_0 + b_1 x_1 + b_2 x_2 + \mathrm{e}$$

通径图如图 9-2 所示。则：

$$r_{10} = p_{0.1} + r_{12} p_{0.2} \qquad (9\text{-}44)$$

$$r_{20} = p_{0.2} + r_{21} p_{0.1} \qquad (9\text{-}45)$$

对于（9-45）式可以进行如下通径分析：由 x_1 到 y 有两条通径，第一条是直接通径 $x_1 \rightarrow y$，直接通径系数 $p_{0.1}$ 表示 x_1 对 y 的直接作用；第二条是间接通径 $x_1 \quad x_2 \rightarrow y$，并定义

间接通径系数等于组成该间接通径的通径与相关线系数的乘积：$r_{12}p_{0.2}$，则间接通径系数表示 x_1 通过与其相关的 x_2 对 y 的间接作用。因此，x_1 与 y 的相关系数 r_{10} 剖分为 x_1 对 y 的直接作用与 x_1 通过 x_2 对 y 的间接作用的代数和，即 x_1 与 y 的相关系数 r_{10} 等于 x_1 与 y 间的直接通径系数 $p_{0.1}$ 与间接通径系数 $r_{12}p_{0.2}$ 之和。

对于（9-45）式也可以进行同样的通径分析：由 x_2 到 y 有两条通径，第一条是直接通径 $x_2 \rightarrow y$，第二条是间接通径 $x_2 \quad \overleftrightarrow{x_1} \rightarrow y$。$x_2$ 与 y 的相关系数 r_{20} 剖分为 x_2 对 y 的直接作用与 x_2 通过 x_1 对 y 的间接作用的代数和，即，x_2 与 y 的相关系数 r_{20} 等于 x_2 与 y 的直接通径系数 $p_{0.2}$ 与间接通径系数 $r_{21}p_{0.1}$ 之和。

对于（9-44）与（9-45）式，可以改写为：

$$\begin{cases} p_{0.1} + r_{12}p_{0.2} = r_{10} \\ r_{12}p_{0.1} + p_{0.2} = r_{20} \end{cases} \tag{9-46}$$

（9-46）式为关于通径系数 $p_{0.1}$、$p_{0.2}$ 的正规方程组。如写成矩阵形式为：

$$\begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \begin{bmatrix} p_{0.1} \\ p_{0.2} \end{bmatrix} = \begin{bmatrix} r_{10} \\ r_{20} \end{bmatrix} \tag{9-47}$$

如果相关系数 r_{12}、r_{10}、r_{20} 已计算出，那么可以由（9-46）式或（9-47）式求出通径系数 $p_{0.1}$ 和 $p_{0.2}$。

一般，设相关变量 y 与 x_1、x_2、…、x_m 间存在线性关系，回归方程为：

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_m\mathrm{x_m}$$

或

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_m\mathrm{x_m} + \mathrm{e}$$

其中，y 为依变量，x_1、x_2、…、x_m 为自变量，且两两相关，通径图如图 9-3 所示。则：

$$\begin{aligned} r_{10} &= p_{0.1} + r_{12}p_{0.2} + \cdots + r_{1m}p_{0.m} \\ r_{20} &= r_{21}p_{0.1} + p_{0.2} + \cdots + r_{2m}p_{0.m} \\ &\vdots \quad\quad \vdots \quad\quad \vdots \quad \cdots \quad \vdots \\ r_{\mathrm{mo}} &= r_{m1}p_{0.1} + r_{m2}p_{0.2} + \cdots + p_{0.m} \end{aligned} \tag{9-48}$$

（9-48）式说明，x_i（$i = 1、2、\cdots、m$）与 y 的相关系数 r_{io} 剖分为 x_i 对 y 的直接作用与间接作用的代数和，即 x_i 与 y 的相关系数 r_{io} 等于 x_i 到 y 的直接通径系数与 x_i 通过与其相关的各个 $x_j\,(j = 1、2、\cdots、m;\ j \neq i)$ 对 y 的所有间接通径系数之和。

对于（9-48）式写成关于各通径系数 $p_{0.1}$、$p_{0.2}$、…、$p_{0.m}$ 的正规方程组形式为：

$$\begin{cases} \mathrm{p}_{0.1} + r_{12}p_{0.2} + \cdots + r_{1m}p_{0.m} = r_{10} \\ r_{20}p_{0.1} + p_{0.2} + \cdots + r_{2m}p_{0.m} = r_{20} \\ \quad\vdots \quad\quad \vdots \quad\quad \cdots \quad\quad \vdots \quad\quad \vdots \\ r_{m1}p_{0.1} + r_{m2}p_{0.2} + \cdots + p_{0.m} = r_{m0} \end{cases} \tag{9-49}$$

（9-49）式的矩阵形式为：

$$\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix} \begin{bmatrix} p_{0.1} \\ p_{0.2} \\ \vdots \\ p_{0.m} \end{bmatrix} = \begin{bmatrix} r_{10} \\ r_{20} \\ \vdots \\ r_{m0} \end{bmatrix} \tag{9-50}$$

如果相关系数 r_{ij}、r_{io}（$i, j = 1、2、\cdots、m$）已经算出，那么可以由（9-49）式或（9-50）式求通径系数 $p_{o.i}$。

性质 1 有两个主要用途：一是利用相关系数求通径系数；二是将原因（自变量）与结果（依变量）的相关系数剖分为直接通径系数与间接通径系数的代数和，从而进行原因对结果的直接作用、间接作用的分析。

性质 2　如果依变量 y 与自变量 x_1、x_2 存在线性关系，且 x_1、x_2 彼此相关（见图 9-2），则原因 x_1、x_2 与剩余项 e 对结果 y 的决定系数 $d_{0.1}$、$d_{0.2}$、$d_{0.e}$ 与 x_1 到 y 的通径系数 $p_{0.1}$、x_2 到 y 的通径系数 $p_{0.2}$ 同 x_1、x_2 间的相关系数 r_{12} 乘积的两倍之和为 1，即：

$$d_{0.1}+d_{0.2}+d_{0.e}+2p_{0.1}r_{12}p_{0.2}=1 \tag{9-51}$$

在（9-51）式中，$2p_{0.1}r_{12}p_{0.2}$ 表示两个相关原因 x_1、x_2 共同对结果 y 的相对决定程度，称为相关原因 x_1、x_2 共同对结果 y 的决定系数，记作 $d_{0.12}$，即：$d_{0.12}=2p_{0.1}r_{12}p_{0.2}$。

因此，（9-51）式可以改写为：

$$d_{0.1}+d_{0.2}+d_{0.12}+d_{0.e}=1 \tag{9-52}$$

即当一个结果的两个原因相关时，两个原因对结果的决定系数加上相关原因共同对结果的决定系数与剩余项对结果的决定系数之和等于 1。

根据（9-52）式，可以计算出剩余项对结果的决定系数 $d_{0.e}$ 与通径系数 $p_{0.e}$：

$$d_{0.e}=1-(d_{0.1}+d_{0.2}+d_{0.12}) \tag{9-53}$$

$$p_{0.e}=\sqrt{d_{0.e}} \tag{9-54}$$

一般，如果依变量 y 与自变量 x_1、x_2、…、x_m 存在线性关系，且自变量两两相关（见图 9-3），则原因 x_1、x_2、…、x_m 分别对结果 y 的决定系数与每两个相关原因共同对结果 y 的决定系数以及剩余项对结果 y 的决定系数之和为 1，即：

$$d_{0.1}+d_{0.2}+\cdots+d_{0.m}+d_{0.12}+d_{0.13}+\cdots+d_{0.(m-1)m}+d_{0.e}=1$$

或简写为：

$$\sum_{i=1}^{m}d_{0.i}+\sum_{i<}^{m}\sum_{j}^{m}d_{0.ij}+d_{0.e}=1 \tag{9-55}$$

根据（9-55）式，可以计算出剩余项对结果的决定系数 $d_{0.e}$ 与通径系数 $p_{0.e}$：

$$d_{0.e}=1-(\sum_{i=1}^{m}d_{0.i}+\sum_{i<}^{m}\sum_{j}^{m}d_{0.ij}) \tag{9-56}$$

$$p_{0.e}=\sqrt{d_{0.e}} \tag{9-57}$$

根据（9-56）式的计算结果，如果 $d_{0.e}$ 的绝对值较大，说明可能还有对结果影响较大的原因未被考虑到，这在通径分析时应予以注意。

性质 2 的主要用途是利用各决定系数绝对值的大小来分析各个原因及两两相关原因对结果的相对决定程度，这是通径分析中进行决定程度分析的依据。同时性质 2 也为计算 $d_{0.e}$、$p_{0.e}$ 提供了简便方法。

性质 3　如果依变量 y 与自变量 x_1、x_2、…、x_m 间存在线性关系，且自变量两两间彼此相关（见图 9-3），回归方程为：

$$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_mx_m$$

或

$$y=b_0+b_1x_1+b_2x_2+\cdots+b_mx_m+e$$

则

$$R^2=p_{0.1}r_{10}+p_{0.2}r_{20}+\cdots+p_{0.m}r_{m0} \tag{9-58}$$

简写为：

$$R^2=\sum_{i=1}^{m}p_{0.i}r_{i0}$$

并且

$$R^2=\sum_{i=1}^{m}d_{0.i}+\sum_{i<}^{m}\sum_{j}^{m}d_{0.ij} \tag{9-59}$$

$$d_{0.e} = 1 - R^2 \tag{9-60}$$

$$p_{0.e} = \sqrt{d_{0.e}} = \sqrt{1 - R^2} \tag{9-61}$$

（9-58）、（9-59）、（9-60）、（9-61）式中 R^2 为依变量 y 与自变量 x_1、x_2、…、x_m 的相关指数，也为标准化变量的回归平方和；（9-60）、（9-61）式中的的 $d_{0.e}$ 为标准化变量的离回归平方和。

（9-58）式中的 $p_{0.1}r_{01}$、$p_{0.2}r_{02}$、…、$p_{0.m}r_{om}$ 可以看成为各原因 x_1、x_2、…、x_m 分别对结果 y 的总决定程度，这是因为根据（9-48）式及两相关原因共同对结果的决定系数的计算公式，可以导出如下结果：

$$\begin{aligned}
p_{0.1}r_{10} &= d_{0.1} + \frac{1}{2}d_{0.12} + \frac{1}{2}d_{0.13} + \cdots + \frac{1}{2}d_{0.1m} \\
p_{0.2}r_{20} &= d_{0.2} + \frac{1}{2}d_{0.21} + \frac{1}{2}d_{0.23} + \cdots + \frac{1}{2}d_{0.2m} \\
&\vdots \quad \vdots \quad \vdots \quad \vdots \quad \cdots \quad \vdots \\
p_{0.m}r_{m0} &= d_{0.m} + \frac{1}{2}d_{0.m1} + \frac{1}{2}d_{0.m2} + \cdots + \frac{1}{2}d_{0.m(m-1)}
\end{aligned} \tag{9-62}$$

根据（9-58）式，将 $p_{0.1}r_{10}$、$p_{0.2}r_{20}$、…、$p_{0.m}r_{m0}$ 分别称为自变量 x_1、x_2、…、x_m 对回归方程估测可靠程度 R^2 的总贡献。

性质 3 的主要用途为在通径分析中进行各原因对回归方程估测可靠程度 R^2 的总贡献分析。

性质 4　如果结果 y_1 与结果 y_2 分别与 x_1、x_2、…、x_m 间存在线性关系，即 x_1、x_2、…、x_m 是影响 y_1 与 y_2 的 m 个公共原因，且 x_1、x_2、…、x_m 两两相关，（见图 9-4）。

则
$$r_{y_1y_2} = \sum_{i=1}^{m} p_{y_1 \cdot x_i} p_{y_2 \cdot x_i} + \sum_{i \neq j}^{m} p_{y_1 \cdot x_i} r_{x_i x_j} p_{y_2 \cdot x_j} \tag{9-63}$$

图 9-4　x_1、x_2、…、x_m 与 y_1 和 y_2 的通径图　　图 9-5　x_1、x_2 与 y_1、y_2 的通径图

性质 4 是通径分析中最重要的性质，利用它，可以计算任意两个结果间的相关系数。性质 4 在动物遗传育种的理论研究上有着十分重要的应用。

例如，结果 y_1 和结果 y_2 分别与 x_1、x_2 存在线性关系，x_1、x_2 是影响 y_1 与 y_2 的两个公共原因，且 x_1 与 x_2 相关（见图 9-5），那么，结果 y_1 与结果 y_2 的相关系数为：

$$r_{y_1y_2} = p_{y_1 \cdot x_1} p_{y_2 \cdot x_1} + p_{y_1 \cdot x_2} p_{y_2 \cdot x_2} + p_{y_1 \cdot x_1} r_{x_1x_2} p_{y_2 \cdot x_2} + p_{y_1 \cdot x_2} r_{x_1x_2} p_{y_2 \cdot x_1}$$

三、通径分析的显著性检验

通径分析是标准化变量的多元线性回归分析。通径分析的显著性检验包括回归方程显著性检验、通径系数显著性检验、通径系数差异显著性检验、两次通径分析相应通径系数差异显著性检验四个内容。后两种检验在一般的多元线性回归分析中通常无法进行，因为偏回归系数所带的单位通常不相同，后两种检验只能在通径分析中进行，这是通径分析的优点之一。

设依变量 y 与自变量 x_1、x_2、…、x_m 间存在线性关系，自变量两两相关。y 与 x_1、

x_2、…、x_m 共有 n 组实际观测数据。则 m 元线性回归方程为：

$$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_mx_{\mathrm{m}} \tag{9-64}$$

现对 y 、 x_1 、 x_2、…、x_m 分别进行标准化变换：

$$y'=\frac{y-\bar{y}}{\sqrt{SS_0}}\text{，}\quad x_i'=\frac{x_i-\bar{x}_i}{\sqrt{SS_i}}\qquad（i=1、2、\cdots、m）$$

其中 $SS_0=\sum(y-\bar{y})^2, SS_i=\sum(x_i-\bar{x}_i)^2$ 。

注意，为了进行显著性检验简便易行，这里我们用 $\sqrt{SS_0}$、$\sqrt{SS_i}$ 作分母进行标准化变换，与前面用 S_0、S_i 作分母进行标准化变换有所不同。两种标准化变换所得的通径系数，即标准化回归变量的偏回归系数数值相同。

标准化变量的 m 元线性回归方程为：

$$\hat{y}'=b_1'x_1'+b_2'x_2'+\cdots+b_m'x_{\mathrm{m}}' \tag{9-65}$$

（9-65）式中，$b_i'=b_i\dfrac{\sqrt{SS_i}}{\sqrt{SS_0}}=b_i\dfrac{s_i}{s_0}$（$i=1、2、\cdots、m$) 为标准化变量的偏回归系数即通径系数。

令 $p_{0.i}=b_i'$，则式（9-65）可改写为：

$$\hat{y}'=p_{0.1}x_1'+p_{0.2}x_2'+\cdots+p_{0.m}x_{\mathrm{m}}' \tag{9-66}$$

（9-66）式也为标准化变量的 m 元线性回归方程。关于通径系数 $p_{0.1}$、$p_{0.2}$、…、$p_{0.m}$ 的正规方程组为：

$$\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix}\begin{bmatrix} p_{0.1} \\ p_{0.2} \\ \vdots \\ p_{0.m} \end{bmatrix}=\begin{bmatrix} r_{0.1} \\ r_{0.2} \\ \vdots \\ r_{0.m} \end{bmatrix}$$

（一）回归方程显著性检验 由于进行通径分析的前提是 y 与 x_1 、 x_2、…、x_m 间存在显著的线性关系，因此需要对回归方程即式（9-66）进行显著性检验。如果经显著性检验，回归方程不显著，则不必继续进行通径分析。

采用 F 检验法检验回归方程的显著性。对于标准化变量的回归方程（9-66）式，总平方和 $\widetilde{ss}_y$ 可以剖分为回归平方和 $\widetilde{SS}_R$ 与离回归平方和 $\widetilde{SS}_r$ 两部分，即：

$$\widetilde{ss}_y=\widetilde{ss}_R+\widetilde{ss}_r \tag{9-67}$$

总自由度 $\widetilde{df}_y$ 可剖分为回归自由度 $\widetilde{df}_R$ 与离回归自由度 $\widetilde{df}_r$ 两部分，即：

$$\widetilde{df}_y=\widetilde{df}_R+\widetilde{df}_r \tag{9-68}$$

可以证明在这里所进行的标准化变化下：

$$\widetilde{ss}_y=1\text{，}\quad \widetilde{df}_y=n-1$$

$$\widetilde{ss}_R=\sum_{i=1}^{m}p_{0.i}r_{i0}=R^2\text{，}\quad \widetilde{df}_R=m$$

$$\widetilde{ss}_r=1-\widetilde{ss}_R=1-R^2\text{，}\quad \widetilde{df}_r=n-m-1$$

由统计量

$$F=\frac{\widetilde{ss}_R/m}{\widetilde{ss}_r/(n-m-1)},(df_1=m,df_2=n-m-1) \tag{9-69}$$

或

$$F=\frac{R^2/m}{(1-R^2)/(n-m-1)},(df_1=m,df_2=n-m-1) \tag{9-70}$$

检验回归方程（9-66）是否显著。

（二）通径系数的显著性检验

1、F检验　由统计量

$$F=\frac{p_{0.i}{}^{2}/c_{ii}}{S\tilde{S}_r/(n-m-1)},\quad (df_1=1,df_2=n-m-1) \tag{9-71}$$

检验通径系数 $p_{0.i}$（$i=1、2、\cdots、m$) 是否显著。其中 c_{ii} 为相关系数矩阵 R 的逆矩阵 $R^{-1}=C$ 中主对角线上的元素。

2、t 检验　由统计量

$$t=\frac{p_{0.i}}{S_{p_{0.i}}},df=n-m-1 \tag{9-72}$$

检验通径系数 $p_{0.i}(i=1、2、\cdots、m)$ 是否显著。其中 $S_{p_{0.i}}$ 为通径系数标准误，其计算公式为：

$$S_{p_{0.i}}=\sqrt{s\tilde{s}_r/(n-m-1)}\cdot\sqrt{c_{ii}} \tag{9-73}$$

注意，这里介绍的 F 检验与 t 检验等价，在实际进行通径系数显著性检验时，只须任选其一。

（三）通径系数差异显著性检验　由于通径系数为不带单位的相对数，因此可以进行一次通径分析中的两个通径系数差异显著性检验。

1、F检验

由统计量

$$F=\frac{(p_{0.i}-p_{0.j})^2/(c_{ii}+c_{jj}-2c_{ij})}{s\tilde{s}_r/(n-m-1)},(df_1=1,df_2=n-m-1) \tag{9-74}$$

$$(i,j=1、2、\cdots、m;\ i\neq j)$$

检验通径系数 $p_{0.i}$ 与 $p_{0.j}$ 之间的差异是否显著。其中 c_{ii}、c_{jj}、c_{ij} 为相关系数矩阵 R 的逆矩阵 $R^{-1}=C$ 的元素。

2、t 检验

由统计量

$$t=\frac{p_{0.i}-p_{0.j}}{S_{p_{0.i}-p_{0.j}}},\qquad (df=n-m-1) \tag{9-75}$$

$$(i,j=1、2、\cdots、m;\ i\neq j)$$

检验通径系数 $p_{0.i}$ 与 $p_{0.j}$ 间的差异是否显著。其中 $S_{p_{0.i}-p_{0.j}}$ 称为通径系数差异标准误，其计算公式为：

$$S_{p_{0.i}-p_{0.j}}=\sqrt{s\tilde{s}_r/(n-m-1)}\cdot\sqrt{cc_{ii}+c_{jj}-2c_{ij}} \tag{9-76}$$

注意，这里介绍的 F 检验与 t 检验也是等价的。

（四）两次通径分析相应通径系数差异显著性检验　设两次通径分析有关数据与结果如下：

	自变量个数	观测值组数	通径系数	离回归平方和	离回归均方	高斯乘数
第一次通径分析	m_1	n_1	$p_{0.i(1)}$	$S\tilde{S}_{r(1)}$	$\tilde{M}S_{r(1)}$	$c_{ii(1)}$
第二次通径分析	m_2	n_2	$p_{0.j(2)}$	$S\tilde{S}_{r(2)}$	$\tilde{M}S_{r(2)}$	$c_{jj(2)}$

$$(i=1、2、\cdots、m_1;\ j=1、2、\cdots、m_2)$$

首先进行两次通径分析剩余项方差齐性检验，这里应用两尾 F 检验。由统计量

$$F=\frac{M\tilde{S}_{r(1)}}{M\tilde{S}_{r(2)}}=\frac{S\tilde{S}_{r(1)}/(n_1-m_1-1)}{S\tilde{S}_{r(2)}/(n_2-m_2-1)},\qquad (\mathrm{df}_1=n_1-m_1-1,df_2=n_2-m_2-1) \tag{9-77}$$

检验 $M\tilde{S}_{r(1)}$ 与 $M\tilde{S}_{r(2)}$ 差异是否显著。注意在式（9-77）中，应将较大的均方放在分子。

如果 $M\tilde{S}_{r(1)}$ 与 $M\tilde{S}_{r(2)}$ 差异不显著，那么再用 F 检验或与其等价的 t 检验进行两次通径分析相应通径系数差异显著性检验。

1、F 检验　由统计量

$$F=\frac{[p_{0.i(1)}-p_{0.j(2)}]^2/[c_{ii(1)}+c_{jj(2)}]}{[\tilde{SS}_{r\ (1)}+\tilde{SS}_{r\ (2)}]/[(n_1-m_1-1)+(n_2-m_2-1)]}\quad[\ \mathrm{df}_1=1, df_2=(n_1-m_1-1)+(n_2-m_2-1)] \tag{9-78}$$

检验两次通径分析相应通径系数 $p_{0.i(1)}$ 与 $p_{0.j(2)}$ 差异是否显著。

2、t 检验　由统计量

$$t=\frac{p_{0.i(1)}-p_{0.j(2)}}{S_{p_{0.i(1)}-p_{0.j(2)}}},\qquad[\mathrm{df}=(\mathrm{n}_1-m_1-1)+(n_2-m_2-1)] \tag{9-79}$$

检验两次通径分析相应通径系数 $p_{0.i(1)}$ 与 $p_{0.j(2)}$ 差异是否显著。其中 $S_{p_{0.i(1)}-p_{0.j(2)}}$ 为两次通径分析的通径系数差异标准误，其计算公式为：

$$S_{P_{0\cdot i(1)}-P_{0\cdot j(2)}}=\sqrt{[\tilde{SS}_{r(1)}+\tilde{SS}_{r(2)}]/[(n_1-m_1-1)+(n_2-m_2-1)]}\cdot\sqrt{C_{ii(1)}+C_{jj(2)}} \tag{9-80}$$

四、实　例

现结合一实例介绍通径分析的基本步骤。

【例 9.4】 奶牛第一胎产奶量是奶牛的重要育种目标，由于奶牛的一个产奶周期较长（305 天），如果能从奶牛的初期性状中找到影响奶牛 305 天产奶量的主要因素，这对保证早期选种的准确性，加速奶牛的育种工作有其重要意义。某奶牛场观察记载了 273 头黑白花奶牛的一胎 305 天产奶量（y），最高日产天数（x_1），最高月产（x_2），90 天产奶量（x_3），最高日产（x_4）5 个性状。试进行通径分析。

（一）计算性状间的相关系数　由 273 组实测数据（略）计算得性状间相关系数并进行显著性检验，如表 9-7 所示。

表 9-7　5 个性状间的相关系数

	最高月产 x_2	90 天产奶量 x_3	最高日产 x_4	一胎 305 天产奶量 y
最高日产天数 x_1	0.1320*	0.0903	0.0864	0.2026**
最高月产 x_2		0.9573**	0.9274**	0.7644**
90 天产奶量 x_3			0.9239**	0.7981**
最高日产 x_4				0.7561**

$r_{0.05(271)}=0.120$　　$\mathrm{r}_{0.01(271)}=0.158$

（二）计算各通径系数　关于通径系数 $p_{0.1}$、$p_{0.2}$、$p_{0.3}$、$p_{0.4}$ 的正规方程组为：

$$\begin{cases} P_{0.1}+0.1320P_{0.2}+0.0903P_{0.3}+0.0864P_{0.4}=0.2026 \\ 0.1320P_{0.1}+P_{0.2}+0.9573P_{0.3}+0.9274P_{0.4}=0.7644 \\ 0.0903P_{0.1}+0.9573P_{0.2}+P_{0.3}+0.9239P_{0.4}=0.7981 \\ 0.0864P_{0.1}+0.9274P_{0.2}+0.9239P_{0.3}+P_{0.4}=0.7561 \end{cases}$$

写成矩阵形式，为：

$$\begin{bmatrix} 1 & 0.1320 & 0.0903 & 0.0864 \\ 0.1320 & 1 & 0.9573 & 0.9274 \\ 0.0903 & 0.9573 & 1 & 0.9239 \\ 0.0864 & 0.9274 & 0.9239 & 1 \end{bmatrix} \begin{bmatrix} P_{0.1} \\ P_{0.2} \\ P_{0.3} \\ P_{0.4} \end{bmatrix} = \begin{bmatrix} 0.2026 \\ 0.7644 \\ 0.7981 \\ 0.7561 \end{bmatrix}$$

求得正规方程组系数矩阵即相关系数矩阵 R 的逆矩阵 $R^{-1}=C$ 如下：

$$C=R^{-1}=\begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ c_{31} & c_{32} & c_{33} & c_{34} \\ c_{41} & c_{42} & c_{43} & c_{44} \end{bmatrix} = \begin{bmatrix} 1.0377 & -0.6530 & 0.3738 & 0.1706 \\ -0.6530 & 14.4930 & -9.8975 & -4.2407 \\ 0.3738 & -9.8975 & 13.5934 & -3.4124 \\ 0.1706 & -4.2407 & -3.4124 & 8.0710 \end{bmatrix}$$

于是求得各通径系数为：

$$\begin{bmatrix} P_{0.1} \\ P_{0.2} \\ P_{0.3} \\ P_{0.4} \end{bmatrix} = \begin{bmatrix} 1.0377 & -0.6530 & 0.3738 & 0.1706 \\ -0.6530 & 14.4930 & -9.8975 & -4.2407 \\ 0.3738 & -9.8975 & 13.5934 & -3.4124 \\ 0.1706 & -4.2407 & -3.4124 & 8.0710 \end{bmatrix} \begin{bmatrix} 0.2026 \\ 0.7644 \\ 0.7981 \\ 0.7561 \end{bmatrix} = \begin{bmatrix} 0.1384 \\ -0.1590 \\ 0.7791 \\ 0.1719 \end{bmatrix}$$

即：$P_{0.1}=0.1384$，$P_{0.2}=-0.1590$，$P_{0.3}=0.7791$，$P_{0.4}=0.1719$

（三）显著性检验

1、线性关系显著性检验——F 检验

因为

$$\widetilde{SS}_R = p_{0.1}r_{10}+p_{0.2}r_{20}+p_{0.3}r_{30}+p_{0.4}r_{40}$$

$$=0.1384\times0.2026+(-0.1590)\times0.7644+0.7791\times0.7981+0.1719\times0.7561$$

$$=0.6583$$

$$\widetilde{SS}_r=1-\widetilde{SS}_R=1-0.6583=0.3417$$

而 $df_R=m=4, df_r=n-m-1=273-4-1=268$

所以 $$F=\frac{\widetilde{SS}_R/m}{\widetilde{SS}_r/(n-m-1)}=\frac{0.6583/4}{0.3417/268}=130.0417^{**}$$

因为 F=130.0417>$F_{0.01(4,268)}$=3.389,P<0.01,表明一胎 305 天产奶量 y 与最高日产天数 x_1、最高月产 x_2、90 天产奶量 x_3、最高日产 x_4 间存在极显著的线性关系,可以对 y 与 x_1、 x_2、 x_3、 x_4 进行通径分析。

又因为 $d_{0.e}=1-R^2=1-\widetilde{SS}_R=\widetilde{SS}_r=0.3417$，所以

$$p_{0.e}=\sqrt{d_{0.e}}=\sqrt{0.3417}=0.5846$$

2、 通径系数显著性检验　选用 F 检验，根据（9-71）式，分别计算检验通径系数 $p_{0.1}$、 $p_{0.2}$、 $p_{0.3}$、 $p_{0.4}$ 显著性的 F 统计量的值，得：

$$F_1=\frac{{p_{0.1}}^2/c_{11}}{S\tilde{S}_r/(n-m-1)}=\frac{0.1384^2/1.0377}{0.3417/(273-4-1)}=14.2308$$

$$F_2=\frac{{p_{0.2}}^2/c_{22}}{SS_r/(n-m-1)}=\frac{(-0.1590)^2/14.4930}{0.3417/(273-4-1)}=1.3077$$

$$F_3=\frac{{p_{0.3}}^2/c_{33}}{S\tilde{S}_r/(n-m-1)}=\frac{0.7791^2/13.5934}{0.3417/(273-4-1)}=34.3846^{**}$$

$$F_4=\frac{{p_{0.4}}^2/c_{44}}{S\tilde{S}_r/(n-m-1)}=\frac{0.1719^2/8.0710}{0.3417/(273-4-1)}=2.8462$$

由 $df_1=1, df_2=n-m-1=268$ 应用线性插值法求得临界 F 值：$F_{0.05(1,268)}=3.88, F_{0.01(1,268)}=6.74$，因 $F_1>F_{0.01(1,268)}$、$F_3>F_{0.01(1,268)}$，而 $F_2<F_{0.05(1,268)}$、$F_4<F_{0.05(1,268)}$，表明通径系数 $P_{0.1}$ 和 $p_{0.3}$ 为极显著，而 $P_{0.2}$ 和 $p_{0.4}$ 为不显著。

3、通径系数差异显著性检验　选用 t 检验法检验 4 个通径系数两两之间的差异显著性。先根据（9—76）式分别算得各通径系数差异标准误如下：

$$\begin{aligned}S_{p_{0.1}-p_{0.2}}&=\sqrt{S\tilde{S}_r/(n-m-1)}\cdot\sqrt{c_{11}+c_{22}-2c_{12}}\\&=\sqrt{0.3417/(273-4-1)}\cdot\sqrt{1.0377+14.4930-2\times(-0.6530)}\\&=0.1465\\S_{p_{0.1}-p_{0.3}}&=\sqrt{S\tilde{S}_r/(n-m-1)}\cdot\sqrt{c_{11}+c_{33}-2c_{13}}\\&=\sqrt{0.3417/(273-4-1)}\cdot\sqrt{1.0377+13.5934-2\times0.3738}\\&=0.1330\\S_{p_{0.1}-p_{0.4}}&=\sqrt{S\tilde{S}_r/(n-m-1)}\cdot\sqrt{c_{11}+c_{44}-2c_{14}}\\&=\sqrt{0.3417/(273-4-1)}\cdot\sqrt{1.0377+8.0710-2\times0.1706}\\&=0.1057\\S_{p_{0.2}-p_{0.3}}&=\sqrt{S\tilde{S}_r/(n-m-1)}\cdot\sqrt{c_{22}+c_{33}-2c_{23}}\\&=\sqrt{0.3417/(273-4-1)}\cdot\sqrt{14.4930+13.5934-2\times(-9.8975)}\\&=0.2471\\S_{p_{0.2}-p_{0.4}}&=\sqrt{S\tilde{S}_r/(n-m-1)}\cdot\sqrt{c_{22}+c_{44}-2c_{24}}\\&=\sqrt{0.3417/(273-4-1)}\cdot\sqrt{14.4930+8.0710-2\times(-4.2407)}\\&=0.1990\\S_{p_{0.3}-p_{0.4}}&=\sqrt{S\tilde{S}_r/(n-m-1)}\cdot\sqrt{c_{33}+c_{44}-2c_{34}}\\&=\sqrt{0.3417/(273-4-1)}\cdot\sqrt{13.5934+8.0710-2\times(-3.4124)}\\&=0.1906\end{aligned}$$

根据（9-75）式算得各 t 统计量的值如下：

$$t_{12}=\frac{p_{0.1}-p_{0.2}}{S_{p_{0.1}-p_{0.2}}}=\frac{0.1384-(-0.1590)}{0.1465}=2.0300^{*}$$

$$t_{13}=\frac{p_{0.1}-p_{0.3}}{S_{p_{0.1}-p_{0.3}}}=\frac{0.1384-0.7791}{0.1330}=-4.8173^{**}$$

$$t_{14}=\frac{p_{0.1}-p_{0.4}}{S_{p_{0.1}-p_{0.4}}}=\frac{0.1384-0.1719}{0.1057}=-0.3169$$

$$t_{23}=\frac{p_{0.2}-p_{0.3}}{S_{p_{0.2}-p_{0.3}}}=\frac{(-0.1590)-0.7791}{0.2471}=-3.7964^{**}$$

$$t_{24}=\frac{p_{0.2}-p_{0.4}}{S_{p_{0.2}-p_{0.4}}}=\frac{(-0.1590)-0.1719}{0.1990}=-1.6628$$

$$t_{34}=\frac{p_{0.3}-p_{0.4}}{S_{p_{0.3}-p_{0.4}}}=\frac{0.7791-0.1719}{0.1906}=3.1857^{**}$$

由 $df=n-m-1=268$ 应用线性内插法求得临界 t 值：$t_{0.05(268)}=1.9704$，$t_{0.01(268)}=2.5975$。因为 $t_{0.05(268)}<t_{12}<t_{0.01(268)}$，表明 $p_{0.1}$ 与 $p_{0.2}$ 差异显著；$|t_{13}|$、$|t_{23}|$、t_{34} 均大于 $t_{0.01(268)}$，表明 $p_{0.3}$ 与 $p_{0.1}$、$p_{0.2}$、$p_{0.4}$ 差异都极显著；而 $|t_{14}|$、$|t_{24}|$ 都小于 $t_{0.05(268)}$，表明 $p_{0.4}$ 与 $p_{0.1}$、$p_{0.2}$ 差异都不显著。

（四）绘制通径图（见图 9-6）　　图 9-6　通径图

（五）进行原因对结果的直接作用与间接作用分析　根据性质 1，可以将自变量（原因）x_1、x_2、x_3、x_4 与依变量（结果）y 的相关系数剖分为直接作用与间接作用的代数和，结果见表 9—8。

表 9-8　直接作用与间接作用分析

性状	相关系数 r_{i0}	直接作用 $p_{0.i}$	间接作用				
			总的	其中通过			
				x_1	x_2	x_3	x_4
x_1	0.2026	0.1384	0.0643		-0.0210	0.0704	0.0149
x_2	0.7644	-0.1590	0.9235	0.0183		0.7548	0.1594
x_3	0.7981	0.7791	0.0191	0.0125	-0.1522		0.1588
x_4	0.7561	0.1719	0.5843	0.0120	-0.1475	0.7198	

（六）进行决定程度分析　各决定系数为:

$$d_{0.1}=p_{0.1}{}^2=0.1384^2=0.0192$$

$$d_{0.2}=p_{0.2}{}^2=(-0.1590)^2=0.0253$$

$$d_{0.3}=p_{0.3}{}^2=0.7791^2=0.6070$$

$$d_{0.4}=p_{0.4}{}^2=0.1719^2=0.0295$$

$$d_{0.e}=S\tilde{S}_r=0.3417$$

$$d_{0.12}=2p_{0.1}r_{12}p_{0.2}=2\times0.1384\times0.1320\times(-0.1590)=-0.0058$$

$$d_{0.13}=2p_{0.1}r_{13}p_{0.3}=2\times0.1384\times0.0903\times0.7791=0.0195$$

$$d_{0.14}=2p_{0.1}r_{14}p_{0.4}=2\times0.1384\times0.0864\times0.1719=0.0041$$

$$d_{0.23}=2p_{0.2}r_{23}p_{0.3}=2\times(-0.1590)\times0.9573\times0.7791=-0.2372$$

$$d_{0.24}=2p_{0.2}r_{24}p_{0.4}=2\times(-0.1590)\times0.9274\times0.1719=-0.0507$$

$$d_{0.34}=2p_{0.3}r_{34}p_{0.4}=2\times0.7791\times0.9239\times0.1719=0.2475$$

按绝对值大小将决定系数进行排列:

$d_{0.3}=0.6070$，　$d_{0.e}=0.3417$，　$d_{0.34}=0.2475$，　$d_{0.23}=-0.2372$，

$d_{0.24}=-0.0507$，　$d_{0.4}=0.0295$，　$d_{0.2}=0.0253$，　$d_{0.13}=0.0195$

$d_{0.1}=0.0192$，　$d_{0.12}=-0.0058$，　$d_{0.14}=0.0041$。

结果表明，x_3 对 y 的相对决定程度最大；其次是剩余项，可能是观测值误差较大或者

可能还有对一胎 305 天产奶量影响较大的性状或因素未考虑到。

（七）进行各自变量对回归方程估测可靠程度 R^2 总贡献分析

先计算各 $P_{0.i}r_{i0}$：

$$P_{0.1}r_{10} = 0.1384\times 0.2026 = 0.0280$$
$$P_{0.2}r_{20} = (-0.1590)\times 0.7644 = -0.1215$$
$$P_{0.3}r_{30} = 0.7791\times 0.7981 = 0.6218$$
$$P_{0.4}r_{40} = 0.1719\times 0.7561 = 0.1300$$

$|p_{0.3}r_{30}| > |p_{0.4}r_{40}| > |p_{0.2}r_{20}| > |p_{0.1}r_{10}|$，说明自变量 x_3 对 R^2 的总贡献为 0.6218，居各自变量对 R^2 总贡献之首。

由上述分析，可以得出如下结论：

（1）一胎 305 天产奶量 y 与最高日产天数 x_1、最高月产 x_2、90 天产奶量 x_3、最高日产 x_4 间存在极显著的线性关系，相关指数 $R^2 = S\tilde{S}_R = 0.6583$，若用 y 与 x_1、x_2、x_3、x_4 间的线性回归方程来估测 y，其可靠程度仅为 65.83%；而剩余项对一胎 305 天产奶量 y 的相对决定程度为 0.3417，其绝对值在各决定系数中居第二，表明可能是观测值误差较大或者可能还有对一胎 305 天产奶量影响较大的性状或因素未被考虑到。

（2）x_1、x_2、x_3、x_4 对 y 的直接作用分别为：$p_{0.1} = 0.1384$，$p_{0.2} = -0.1590$，$p_{0.3} = 0.7791$，$p_{0.4} = 0.1719$。其中 $p_{0.2}$、$p_{0.4}$ 不显著，$p_{0.1}$、$p_{0.3}$ 达极显著。但 $p_{0.3}$ 极显著地高于 $p_{0.1}$、$p_{0.2}$、$p_{0.4}$，表明，4 个性状中最高日产天数 x_1 和 90 天产奶量 x_3 对一胎 305 天产奶量 y 有极显著影响，而 90 天产奶量 x_3 又极显著地高于其它 3 个性状对一胎 305 天产奶量 y 的影响，由此说明 90 天产奶量是影响一胎 305 天产奶量的最重要的早期性状。

（3）从表 9-8 看到，x_3 对 y 的直接作用最大，$p_{0.3}$=0.7791，x_3 通过 x_1、x_2、x_4 对 y 的间接作用之和为 0.0191，其绝对值较小。x_1 也具有类似的特性。表明 x_3、x_1 对 y 的作用主要为直接作用，而 x_2、x_4 对 y 的作用主要为间接作用。

（4）90 天产奶量 x_3 与最高日产 x_4 共同对一胎 305 天产奶量 y 的相对决定程度为 0.2475，其绝对值在各决定系数中居第三；且 x_4 对 R^2 的总贡献中居第二，说明在注意 90 天产奶量 x_3 的同时，还应注意最高日产 x_4 这一早期产奶性状，若二者皆高，则一胎 305 天产奶量很可能也是高的，由于 $r_{34} = 0.9239^{**}$，x_3 与 x_4 为极显著正相关，两个性状同步增减，易于实现两个性状都高。

（5）90 天产奶量 x_3 与最高月产 x_2 共同对一胎 305 天产奶量 y 的相对决定程度为 -0.2372，之所以是负的，是因为 $p_{0.2} = -0.1590$，即 x_2 对 y 的直接作是负的。但是 $r_{23} = 0.9573^{**}$，即 x_2 与 x_3 为极显著正相关，x_3 高（这是我们所希望的），一般说来 x_2 也高（这是我们所不希望的）。所以当注意了 90 天产奶量 x_3 这一性状后，还应尽量兼顾最高月产 x_2 不要太高，否则会影响一胎 305 天产奶量。

因此，为了选取一胎 305 天产奶量高的奶牛，就所考虑的 4 个早期性状而言，应选取 90 天产奶量高、最高日产量高而最高月产奶量适中的奶牛；此外，还应进一步寻找对奶牛一胎 305 天产奶量影响较大的另外的性状或因素。

习　　题

1.如何建立多元线性回归方程？偏回归系数有何意义？

2．多元线性回归的显著性检验包含哪些内容？如何进行？

3．在多元线性回归分析中，如何剔除不显著的自变量？怎样重新建立多元线性回归方程？

4．什么是相关系数？其意义是什么？

5．什么是复相关系数？其意义是什么？如何进行显著性检验？

6．什么是偏相关系数？偏相关分析与简单相关分析有何区别？

7. 如何将多项式回归转化为多元线性回归？

8．什么是通径系数？怎样计算和检验通径系数？通径分析的基本步骤有哪些？

9．根据下述某猪场 25 头育肥猪 4 个胴体性状的数据资料，试进行瘦肉量 y 对眼肌面积（x_1）、腿肉量(x_2)、腰肉量(x_3)的多元线性回归分析。

序号	瘦肉量 $y(kg)$	眼肌面积 $x_1(cm^2)$	腿肉量 $x_2(kg)$	腰肉量 $x_3(kg)$	序号	瘦肉量 $y(kg)$	眼肌面积 $x_1(cm^2)$	腿肉量 $x_2(kg)$	腰肉量 $x_3(kg)$
1	15.02	23.73	5.49	1.21	14	15.94	23.52	5.18	1.98
2	12.62	22.34	4.32	1.35	15	14.33	21.86	4.86	1.59
3	14.86	28.84	5.04	1.92	16	15.11	28.95	5.18	1.37
4	13.98	27.67	4.72	1.49	17	13.81	24.53	4.88	1.39
5	15.91	20.83	5.35	1.56	18	15.58	27.65	5.02	1.66
6	12.47	22.27	4.27	1.50	19	15.85	27.29	5.55	1.70
7	15.80	27.57	5.25	1.85	20	15.28	29.07	5.26	1.82
8	14.32	28.01	4.62	1.51	21	16.40	32.47	5.18	1.75
9	13.76	24.79	4.42	1.46	22	15.02	29.65	5.08	1.70
10	15.18	28.96	5.30	1.66	23	15.73	22.11	4.90	1.81
11	14.20	25.77	4.87	1.64	24	14.75	22.43	4.65	1.82
12	17.07	23.17	5.80	1.90	25	14.37	20.44	5.10	1.55
13	15.40	28.57	5.22	1.66					

($\hat{y}=0.8563+0.0187x_1+2.0729x_2+1.9380x_3$；　$F=37.1560**, F_{b_1}=0.4002, F_{b_2}=58.8795**,$
$F_{b_3}=14.2513**; \hat{y}=1.1286+2.1019x_2+1.9764x_3; F=57.0842**, F_{b_2'}=64.0778**,$
$F_{b_3'}=15.4508**$)

10.对本章习题 9 所给数据资料,分别计算 y 与 x_1、 x_2 、 x_3 的二级偏相关系数并进行显著性检验。

（$r_{01.23}=0.1366, r_{02.13}=0.8585**, r_{03.12}=0.6539**$）

11.根据重庆市种畜场奶牛群各月份产犊母牛平均 305 天产奶量的数据资料，试进行一元二次多项式回归分析。

平均产奶量 y（kg）	3833.43	3811.58	3769.47	3565.74	3481.99	3372.82
产犊月份 x	1	2	3	4	5	6
平均产奶量 y（kg）	3476.76	3466.22	3395.42	3807.08	3817.03	3884.52
产犊月份 x	7	8	9	10	11	12

($\hat{y}=4117.14-204.9362x+15.7857x^2; F=16.8923**$)

12．根据【例 9.1】所给数据资料，试进行瘦肉量 y 与眼肌面积（ x_1 ）、胴体长（ x_2 ）、膘厚（ x_3 ）四个性状的通径分析。

（ $p_{0.1}=0.4436, p_{0.2}=0.3323, p_{0.3}=0.2459$ ）

第十章　协方差分析

第一节　协方差分析的意义

协方差分析有二个意义，一是对试验进行统计控制，二是对协方差组分进行估计，现分述如下。

一、对试验进行统计控制

为了提高试验的精确性和准确性，对处理以外的一切条件都需要采取有效措施严加控制，使它们在各处理间尽量一致，这叫试验控制。但在有些情况下，即使作出很大努力也难以使试验控制达到预期目的。例如：研究几种配合饲料对猪的增重效果，希望试验仔猪的初始重相同，因为仔猪的初始重不同，将影响到猪的增重。经研究发现：增重与初始重之间存在线性回归关系。但是，在实际试验中很难满足试验仔猪初始重相同这一要求。这时可利用仔猪的初始重(记为x)与其增重(记为y)的回归关系，将仔猪增重都矫正为初始重相同时的增重，于是初始重不同对仔猪增重的影响就消除了。由于矫正后的增重是应用统计方法将初始重控制一致而得到的，故叫统计控制。统计控制是试验控制的一种辅助手段。经过这种矫正，试验误差将减小，对试验处理效应估计更为准确。若y的变异主要由x的不同造成(处理没有显著效应)，则各矫正后的y'间将没有显著差异(但原y间的差异可能是显著的)。若y的变异除掉x不同的影响外，尚存在不同处理的显著效应，则可期望各y'间将有显著差异(但原y间差异可能是不显著的)。此外，矫正后的y'和原y的大小次序也常不一致。所以，处理平均数的回归矫正和矫正平均数的显著性检验，能够提高试验的准确性和精确性，从而更真实地反映试验实际。这种将回归分析与方差分析结合在一起，对试验数据进行分析的方法，叫做协方差分析(**analysis of covariance**)。

二、估计协方差组分

在第八章曾介绍过表示两个相关变量线性相关性质与程度的相关系数的计算公式：

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

若将公式右端的分子分母同除以自由度(n–1)，得

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})/(n-1)}{\sqrt{\left[\sum(x-\bar{x})^2\Big/(n-1)\right]\left[\sum(y-\bar{y})^2\Big/(n-1)\right]}} \qquad (10\text{–}1)$$

其中

$\frac{\sum(x-\bar{x})^2}{n-1}$ 是x的均方MS_x，它是x的方差 σ_x^2 的无偏估计量；

$\frac{\sum(y-\bar{y})^2}{n-1}$ 是y的均方MS_y，它是y的方差 σ_y^2 的无偏估计量；

$\frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$称为$x$与$y$的平均的离均差的乘积和，简称均积，记为$MP_{xy}$，即

$$MP_{xy}=\frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}=\frac{\sum xy-\frac{(\sum x)(\sum y)}{n}}{n-1} \qquad (10\text{-}2)$$

与均积相应的总体参数叫协方差（**covariance**），记为$COV(x,y)$或σ_{xy}。统计学证明了，均积MP_{xy}是总体协方差$COV(x,y)$的无偏估计量，即 $EMP_{xy}= COV(x,y)$。

于是，样本相关系数r可用均方MS_x、MS_y，均积MP_{xy}表示为：

$$r=\frac{MP_{xy}}{\sqrt{MS_x MS_y}} \qquad (10\text{-}3)$$

相应的总体相关系数ρ可用x与y的总体标准差σ_x、σ_y，总体协方差COV(x,y)或σ_{xy}表示如下：

$$\rho=\frac{COV(x,y)}{\sigma_x\sigma_y}=\frac{\sigma_{xy}}{\sigma_x\sigma_y} \qquad (10\text{-}4)$$

均积与均方具有相似的形式，也有相似的性质。在方差分析中，一个变量的总平方和与自由度可按变异来源进行剖分，从而求得相应的均方。统计学已证明：两个变量的总乘积和与自由度也可按变异来源进行剖分而获得相应的均积。这种把两个变量的总乘积和与自由度按变异来源进行剖分并获得获得相应均积的方法亦称为协方差分析。

在随机模型的方差分析中，根据均方MS和期望均方EMS的关系，可以得到不同变异来源的方差组分的估计值。同样，在随机模型的协方差分析中，根据均积MP和期望均积EMP的关系，可得到不同变异来源的协方差组分的估计值。有了这些估计值，就可进行相应的总体相关分析。这些分析在遗传、育种和生态、环保的研究上是很有用处的。

由于篇幅限制，本章只介绍对试验进行统控制的协方差分析。

第二节　单因素试验资料的协方差分析

设有k个处理、n次重复的双变量试验资料，每处理组内皆有n对观测值x、y，则该资料为具kn对x、y观测值的单向分组资料，其数据一般模式如表10—1所示。

表10—1　kn对观测值x、y的单向分组资料的一般形式

处　理	处理1		处理2		…	处理i		…	处理k	
观测指标	x	y	x	y	…	x	y	…	x	y
观测值	x_{11}	y_{11}	x_{21}	y_{21}	…	x_{i1}	y_{i1}		x_{k1}	y_{k1}
	x_{12}	y_{12}	x_{22}	y_{22}	…	x_{i2}	y_{i2}	…	x_{k2}	y_{k2}
x_{ij}、y_{ij}	…	…	…	…	…	…	…	…	…	…
(i=1,2,…k	x_{1j}	y_{1j}	x_{2j}	y_{2j}	…	x_{ij}	y_{ij}	…	x_{kj}	y_{kj}
j=1,2,…n)	…	…	…	…	…	…	…	…	…	…
	x_{1n}	y_{1n}	x_{2n}	y_{2n}	…	x_{in}	y_{in}	…	x_{kn}	y_{kn}
总　和	$x_{1\cdot}$	$y_{1\cdot}$	$x_{2\cdot}$	$y_{2\cdot}$	…	$x_{i\cdot}$	$y_{i\cdot}$	…	$x_{k\cdot}$	$y_{k\cdot}$
平均数	$\bar{x}_1$.	$\bar{y}_1$.	$\bar{x}_2$.	$\bar{y}_2$.	…	$\bar{x}_i$.	$\bar{y}_i$.	…	$\bar{x}_k$.	$\bar{y}_k$.

表10—1的x和y变量的自由度和平方和的剖分参见单因素试验资料的方差分析方法一

节。其乘积和的剖分则为：

总变异的乘积和SP_T是x_{ij}与$\bar{x}..$和y_{ij}与$\bar{y}..$的离均差乘积之和，即：

$$SP_T=\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}..)(y_{ij}-\bar{y}..)=\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij}y_{ij}-\frac{x..y..}{kn} \tag{10-5}$$

$$df_T=kn\text{-}1$$

其中，$x..=\sum_{i=1}^{k}x_i.$，$y..=\sum_{i=1}^{k}y_i.$，$\bar{x}..=x..\big/kn$，$\bar{y}..=y..\big/kn$ 。

处理间的乘积和SP_t是$\bar{x}_i.$与$\bar{x}..$和$\bar{y}_i.$与$\bar{y}..$的离均差乘积之和乘以n，即：

$$SP_t=n\sum_{i=1}^{k}(\bar{x}_i.-\bar{x}..)(\bar{y}_i.-\bar{y}..)=\frac{1}{n}\sum_{i=1}^{k}x_i.y_i.-\frac{x_i.y_i.}{kn} \tag{10-6}$$

$$df_t=k-1$$

处理内的乘积和SP_e是x_{ij}与$\bar{x}_i.$和y_{ij}与$\bar{y}_i.$的离均差乘积之和，即：

$$SP_e=\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i.)(y_{ij}-\bar{y}_i.)=\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij}y_{ij}-\frac{1}{n}\sum_{i=1}^{k}x_i.y_i.=SP_T-SP_t \tag{10-7}$$

$$df_e=k(n-1)$$

以上是各处理重复数n相等时的计算公式，若各处理重复数n不相等，分别为n_1、n_2、…、n_k，其和为$\sum_{i=1}^{k}n_i$，则各项乘积和与自由度的计算公式为：

$$SP_T=\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}y_{ij}-\frac{x_i.y_i.}{\sum_{i=1}^{k}n_i}$$

$$df_T=\sum_{i=1}^{k}n_i-1 \tag{10-8}$$

$$SP_t=\frac{x_1.y_1.}{n_1}+\frac{x_2.y_2.}{n_2}+\ldots+\frac{x_k.y_k.}{n_k}-\frac{x..y..}{\sum_{i=1}^{k}n_i}$$

$$df_t=k-1$$

$$SP_e=\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}y_{ij}-\left[\frac{x_1.y_1.}{n_1}+\frac{x_2.y_2.}{n_2}+\ldots+\frac{x_k.y_k.}{n_k}\right]=SP_T-SP_t$$

$$df_e=\sum_{i=1}^{k}n_i-k=df_T-df_t \tag{10-9}$$

有了上述SP和df，再加上x和y的相应SS，就可进行协方差分析。

【例10.1】 为了寻找一种较好的哺乳仔猪食欲增进剂，以增进食欲，提高断奶重，对哺乳仔猪做了以下试验：试验设对照、配方1、配方2、配方3共四个处理，重复12次，选择初始条件尽量相近的长白种母猪的哺乳仔猪48头，完全随机分为4组进行试验，结果见表10—2，试作分析。

此例，$x..=x_1.+x_2.+x_3.+x_4.$ =18.25+15.40+15.65+13.85=63.15

$y..=y_1.+y_2.+y_3.+y_4.$=141.80+130.10+144.80+133.80=550.50

k=4，n=12，kn=4×12=48

表10—2　不同食欲增进剂仔猪生长情况表　（单位：*kg*）

处　理	对照		配方1		配方2		配方3	
观　测	初生重	50日	初生重	50日	初生重	50日	初生	50日
指　标	x	龄重y	x	龄重y	x	龄重y	重x	龄重y
	1.50	12.40	1.35	10.20	1.15	10.00	1.20	12.40
	1.85	12.00	1.20	9.40	1.10	10.60	1.00	9.80
	1.35	10.80	1.45	12.20	1.10	10.40	1.15	11.60
	1.45	10.00	1.20	10.30	1.05	9.20	1.10	10.60
	1.40	11.00	1.40	11.30	1.40	13.00	1.00	9.20
观 察 值	1.45	11.80	1.30	11.40	1.45	13.50	1.45	13.90
x_{ij},y_{ij}	1.50	12.50	1.15	12.80	1.30	13.00	1.35	12.80
	1.55	13.40	1.30	10.90	1.70	14.80	1.15	9.30
	1.40	11.20	1.35	11.60	1.40	12.30	1.10	9.60
	1.50	11.60	1.15	8.50	1.45	13.20	1.20	12.40
	1.60	12.60	1.35	12.20	1.25	12.00	1.05	11.20
	1.70	12.50	1.20	9.30	1.30	12.80	1.10	11.00
总 和 $x_{i\cdot}$, $y_{i\cdot}$	18.25	141.80	15.40	130.80	15.65	144.80	13.85	133.80
平 均 $\bar{x}_{i\cdot}$, $\bar{y}_{i\cdot}$	1.52	11.82	1.28	10.84	1.30	12.07	1.15	1.15

协方差分析的计算步骤如下：

(一)求x变量的各项平方和与自由度

1、总平方和及自由度

$$SS_{T(x)}=\sum\sum x_{ij}^2-\frac{x_{\cdot\cdot}^{\ 2}}{kn}=(1.50^2+1.85^2+\ldots+1.10^2)-\frac{63.15^2}{48}=84.8325-\frac{63.15^2}{48}=1.75$$

$df_{T(x)}=kn-1=4\times12-1=47$

2、处理间平方和与自由度

$$SS_{t(x)}=\frac{1}{n}\sum_{i=1}^{k}x_{i\cdot}^2-\frac{x_{\cdot\cdot}^{\ 2}}{kn}=\frac{1}{12}(18.25^2+15.40^2+15.65^2+13.85^2)-\frac{63.15^2}{48}=0.83$$

$df_{t(x)}=k-1=4-1=3$

3、处理内平方和与自由度

$SS_{e(x)}=SS_{T(x)}-SS_{t(x)}=1.75-0.83=0.92$

$df_{e(x)}=df_{T(x)}-df_{t(x)}=47-3=44$

(二)求y变量各项平方和与自由度

1、总平方和与自由度

$$SS_{T(y)}=\sum\sum y_{ij}^2-\frac{y_{\cdot\cdot}^2}{kn}=(12.40^2+12.00^2+\ldots+11.00^2)-\frac{550.5^2}{48}=6410.31-\frac{550.5^2}{48}=96.76$$

$df_{T(y)}=kn-1=4\times12-1=47$

2、处理间平方和与自由度

$$SS_{t(y)}=\frac{1}{n}\sum y_{i\cdot}^2-\frac{y_{\cdot\cdot}^2}{kn}=\frac{1}{12}(141.80^2+130.80^2+144.80^2+133.80^2)-\frac{550.50^2}{48}=11.68$$

$df_{t(y)}=k-1=4-1=3$

3、处理内平方和与自由度

$SS_{e(y)}=SS_{T(y)}-SS_{t(y)}=96.76-11.68=85.08$

$df_{e(y)} = df_{T(y)} - df_{t(y)}$ =47-3=44

(三)求*x*和*y*两变量的各项离均差乘积和与自由度

1、总乘积和与自由度

$$SP_T = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}y_{ij} - \frac{x..y..}{kn}$$

$$= 1.50\times12.40 + 1.85\times12.00 + ... + 1.10\times11.00 - \frac{63.15\times550.50}{4\times12}$$

$$= 732.50 - \frac{63.15\times550.50}{4\times12} = 8.25$$

$df_{T(x,y)}$ =*kn*-1=4×12-1=47

2、处理间乘积和与自由度

$$SP_t = \frac{1}{n}\sum_{i=1}^{k} x_i.y_i. - \frac{x..y..}{kn}$$

$$= \frac{1}{12}(18.25\times141.80 + 15.40\times130.10 + 15.65\times144.80 + 13.85\times133.80) - \frac{63.15\times550.50}{4\times12}$$

=1.64

$df_{t(x,y)}$ =*k*-1=4-1=3

3、处理内乘积和与自由度

$SP_e = SP_T - SP_t$ =8.25-1.64=6.61

$df_{e(x,y)} = df_{T(x,y)} - df_{t(x,y)}$ =47-3=44

平方和、乘积和与自由度的计算结果列于表10—3。

表10—3　*x*与*y*的平方和与乘积和表

变异来源	*df*	SS_x	SS_y	SP_{xy}
处理间(*t*)	3	0.83	11.68	1.64
处理内(误差)(*e*)	44	0.92	85.08	6.61
总变异(*T*)	47	1.75	96.76	8.25

(四)对*x*和*y*各作方差分析(表10—4)

表10—4　初生重与50日龄重的方差分析表

变异来源	*df*	*x*变量			*y*变量			*F*值
		SS	*MS*	*F*	*SS*	*MS*	*F*	
处理间	3	0.83	0.28	13.33**	11.68	3.89	2.02	$F_{0.05}$=2.82
处理内(误差)	44	0.92	0.021		85.08	1.93		$F_{0.01}$=4.26
总变异	47	1.75			96.76			

分析结果表明，4种处理的供试仔猪平均初生重间存在着极显著的差异，其50 日龄平均重差异不显著。须进行协方差分析，以消除初生重不同对试验结果的影响，减小试验误差，揭示出可能被掩盖的处理间差异的显著性。

(五)协方差分析

1、误差项回归关系的分析　误差项回归关系分析的意义是要从剔除处理间差异的影响的误差变异中找出50日龄重(*y*)与初生重(*x*)之间是否存在线性回归关系。计算出误差项的回归系数并对线性回归关系进行显著性检验，若显著则说明两者间存在回归关系。这时就

可应用线性回归关系来校正y值(50日龄重)以消去仔猪初生重(x)不同对它的影响。然后根据校正后的y值(校正50日龄重)来进行方差分析。如线性回归关系不显著，则无需继续进行分析。

回归分析的步骤如下：

(1)计算误差项回归系数，回归平方和，离回归平方和与相应的自由度

从误差项的平方和与乘积和求误差项回归系数：

$$b_{yx(e)}=\frac{SP_e}{SS_{e(x)}}=\frac{6.61}{0.92}=7.1848 \tag{10-10}$$

误差项回归平方和与自由度

$$SS_{R(e)}=\frac{SP_e{}^2}{SS_{e(x)}}=\frac{6.61^2}{0.92}=47.49 \tag{10-11}$$

$df_{R(e)}$=1

误差项离回归平方和与自由度

$$SS_{r(e)}=SS_{e(y)}-SS_{R(e)}=85.08-47.49=37.59 \tag{10-12}$$

$$df_{r(e)}=df_{e(y)}-df_{R(e)}=44-1=43$$

(2)检验回归关系的显著性(表10—5)

表10—5　哺乳仔猪50日龄重与初生重的回归关系显著性检验表

变异来源	*SS*	*df*	*MS*	*F*	$F_{0.01}$
误差回归	47.49	1	47.49	54.32**	7.255
误差离回归	37.59	43	0.8742		
误差总和	85.08	44			

F检验表明，误差项回归关系极显著，表明哺乳仔猪50 日龄重与初生重间存在极显著的线性回归关系。因此，可以利用线性回归关系来校正y，并对校正后的y进行方差分析。

2、对校正后的50日龄重作方差分析

(1)求校正后的50日龄重的各项平方和及自由度　利用线性回归关系对50日龄重作校正，并由校正后的50日龄重计算各项平方和是相当麻烦的，统计学已证明，校正后的总平方和、误差平方和及自由度等于其相应变异项的离回归平方和及自由度，因此，其各项平方和及自由度可直接由下述公式计算。

①校正50日龄重的总平方和与自由度，即总离回归平方和与自由度

$$SS_T^{'}=SS_{T(y)}-SS_{R(y)}=SS_{T(y)}-\frac{SP_T{}^2}{SS_{T(x)}}=96.76-\frac{8.25^2}{1.75}=57.85 \tag{10-13}$$

$$df_T^{'}=df_{T(y)}-df_{R(y)}=47-1=46$$

②校正50日龄重的误差项平方和与自由度，即误差离回归平方和与自由度

$$SS_e^{'}=SS_{e(y)}-SS_{R(e)}=SS_{e(y)}-\frac{SP_e{}^2}{SS_{e(x)}}=85.08-\frac{6.61^2}{0.92}=37.59 \tag{10-14}$$

$$df_e^{'}=df_{e(y)}-df_{e(R)}=44-1=43$$

上述回归自由度均为1，因仅有一个自变量x。

③校正50日龄重的处理间平方和与自由度

$$SS_t'=SS_T'-SS_e'=57.87-37.59=20.28 \tag{10-15}$$

$$df_t'=df_T'-df_e'=k-1=4-1=3$$

(2)列出协方差分析表，对校正后的50日龄重进行方差分析(表10—6)

查F表：$F_{0.01(3,43)}$=4.275(由线性内插法计算)，由于F=7.63＞$F_{0.01(3,43)}$，P＜0.01，表明对于校正后的50日龄重不同食欲添加剂配方间存在极显著的差异。故须进一步检验不同处理间的差异显著性，即进行多重比较。

表10—6　表10-2资料的协方差分析表

变异来源	df	SS_x	SS_y	SP_{xy}	b	校正50日龄重的方差分析			F
						df'	SS'	MS	
处理间(t)	3	0.83	11.68	1.64					
机　误(e)	44	0.92	85.08	6.61	7.1848	43	37.59	0.8742	
总　和(T)	47	1.75	96.76	8.25		46	57.87		
校正处理间						3	20.28	6.76	7.63**

3、根据线性回归关系计算各处理的校正50日龄平均重

误差项的回归系数$b_{yx(e)}$表示初生重对50日龄重影响的性质和程度，且不包含处理间差异的影响，于是可用$b_{yx(e)}$根据平均初生重的不同来校正每一处理的50日龄平均重。校正50日龄平均重计算公式如下：

$$\bar{y}'_{i}.=\bar{y}_{i}.-b_{yx(e)}(\bar{x}_{i}.-\bar{x}..) \tag{10-16}$$

公式中：$\bar{y}'_{i}.$为第i处理校正50日龄平均重；

$\bar{y}_{i}.$为第i处理实际50日龄平均重(见表10—2)；

$\bar{x}_{i}.$为第i处理实际平均初生重(见表10—2)；

$\bar{x}..$为全试验的平均数，$\bar{x}..=\dfrac{x..}{kn}=\dfrac{63.15}{48}=1.3156$

$b_{yx(e)}$为误差回归系数，$b_{yx(e)}$=7.1848

将所需要的各数值代入(10—16)式中，即可计算出各处理的校正50日龄平均重(见表10—7)。

表10—7　各处理的校正50日龄平均重计算表

处　理	$\bar{x}_{i}.-\bar{x}..$	$b_{yx(e)}(\bar{x}_{i}.-\bar{x}..)$	实际50日龄平均重	校正50日龄平均重 $\bar{y}_{i}.-b_{yx(e)}(\bar{x}_{i}.-\bar{x}..)$
对　照	1.52-1.3156=0.2044	7.1848×0.2044=1.4686	11.82	11.82-1.1686=10.3514
配方1	1.28-1.3156=-0.0356	7.1848×（-0.0356）=-0.2588	10.84	10.84+0.2558=12.0758
配方2	1.30-1.3156=-0.0156	7.1848×（-0.0156）=-0.1121	12.07	12.07+0.1121=12.1821
配方3	1.15-1.3156=-0.1656	7.1848×（-0.1656）=-1.1898	11.15	11.15+1.1898=12.3398

4、各处理校正50日龄平均重间的多重比较

各处理校正50日龄平均重间的多重比较，即各种食欲添加剂的效果比较。

(1)t检验　检验两个处理校正平均数间的差异显著性，可应用t检验法：

$$t=\frac{\bar{y}'_{i}.-\bar{y}'_{j}.}{S_{\bar{y}'_{i}.-\bar{y}'_{j}.}} \tag{10-17}$$

$$S_{\bar{y}_{i}.'-\bar{y}_{j}.'}=\sqrt{MS'_{e}\left[\frac{2}{n}+\frac{(\bar{x}_{i}.-\bar{x}_{j}.)^{2}}{SS_{e(x)}}\right]} \tag{10-18}$$

式中，$\bar{y}'_{i}.-\bar{y}'_{j}.$为两个处理校正平均数间的差异；

$S_{\bar{y}_i'-\bar{y}_j'}$为两个处理校正平均数差数标准误；

MS_e'为误差离回归均方；

n为各处理的重复数；

$\bar{x}_i.$为处理i的x变量的平均数；

$\bar{x}_j.$为处理j的x变量的平均数；

$SS_{e(x)}$为x变量的误差平方和

例如，检验食欲添加剂配方1与对照校正50日龄平均重间的差异显著性：

$$\bar{y}_1'.-\bar{y}_2'.=10.3514-12.0758=-1.7244$$

$$MS_e'=37.59/43=0.8742 \qquad n=12$$

$$\bar{x}_1.=1.52,\ \bar{x}_2.=1.28,\ SS_{e(x)}=0.92$$

将上面各数值代入(10—18)式得：

$$S_{\bar{y}_1'-\bar{y}_2'}=\sqrt{0.8742\times\left[\frac{2}{12}+\frac{(1.52-1.28)^2}{0.92}\right]}=0.4477$$

于是 $$t=\frac{10.3514-12.0758}{0.4477}=-3.85$$

查t值表，当自由度为43时(见表10—6误差自由度)，$t_{0.01(43)}$=2.70(利用线性内插法计算)，$|t|>t_{0.01(43)}$，$P<0.01$，表明对照与食欲添加剂1号配方校正50日龄平均重间存在着极显著的差异，这里表现为1号配方的校正50日龄平均重极显著高于对照。 其余的每两处理间的比较都须另行算出$S_{\bar{y}_i'-\bar{y}_j'}$，再进行$t$检验。

（2）最小显著差数法　利用t检验法进行多重比较，每一次比较都要算出各自的$S_{\bar{y}_i'-\bar{y}_j'}$，比较麻烦。当误差项自由度在 20以上，$x$变量的变异不甚大(即$x$变量各处理平均数间差异不显著)，为简便起见，可计算一个平均的$\bar{S}_{\bar{y}_i'-\bar{y}_j'}$采用最小显著差数法进行多重比较。$\bar{S}_{\bar{y}_i'-\bar{y}_j'}$的计算公式如下：

$$\bar{S}_{\bar{y}_i'-\bar{y}_j'}=\sqrt{\frac{2MS_e'}{n}\left[1+\frac{SS_{t(x)}}{SS_{e(x)}(k-1)}\right]} \tag{10-19}$$

公式中$SS_{t(x)}$为x变量的处理间平方和。

然后按误差自由度查临界t值，计算出最小显著差数：

$$LSD_\alpha=t_{\alpha(dfe)}\ \bar{S}_{\bar{y}_i'-\bar{y}_j'} \tag{10-20}$$

本例x变量处理平均数间差异极显著，不满足“x变量的变异不甚大”这一条件，不应采用此处所介绍的最小显著差数法进行多重比较。为了便于读者熟悉该方法，仍以本例的数据说明之。此时

$$\bar{S}_{\bar{y}_i'-\bar{y}_j'}=\sqrt{\frac{2\times0.8742}{12}\left[1+\frac{0.83}{0.92\times(4-1)}\right]}=0.4354$$

由df_e'=43, 查临界t值得：$t_{0.05(43)}$=2. 017，$t_{0.01(43)}$=2. 70

于是 $$LSD_{0.05}=2.017\times0.4353=0.878$$

$$LSD_{0.01}=2.70\times0.4353=1.175$$

不同食欲添加剂配方与对照校正50日龄平均重比较结果见表10—8。

表10—8　不同食欲添加剂配方与对照间的效果比较表

食欲添加剂配方	校正50日龄平均重	对照校正50日龄平均重	差数

1	12.0758	10.3514	1.7244**
2	12.1821	10.3514	1.8307**
3	12.3398	10.3514	1.9884**

多重比较结果表明：食欲添加剂配方1、2、3号与对照比较，其校正50 日龄平均重间均存在极显著的差异，这里表现为配方1、2、3号的校正50日龄平均重均极显著高于对照。

(3)最小显著极差法　　当误差自由度在20以上，x变量的变异不甚大，还可以计算出平均的平均数校正标准误 $\overline{S}_{\bar{y}}$，利用LSR 法进行多重比较。 $\overline{S}_{\bar{y}}$ 的计算公式如下：

$$\overline{S}_{\bar{y}} = \sqrt{\frac{MS_e^{'}}{n}\left[1+\frac{SS_{t(x)}}{SS_{e(x)}(k-1)}\right]} \qquad (10\text{-}21)$$

然后由误差自由度 df_e' 和秩次距k查SSR表（或q表），计算最小显著极差：

$$LSR_{\alpha} = SSR_{\alpha}\overline{S}_{\bar{y}} \qquad (10\text{-}22)$$

对于【例10.1】资料，由于不满足“x变量的变异不甚大”这一条件，不应采用此处所介绍的LSR法进行多重比较。为了便于读者熟悉该方法，仍以【例10.1】的数据说明之。此时

$\mathrm{MS}_{\mathrm{e}}^{'}$=0.8742，$n$=12，$SS_{t(x)}$=0.83，$SS_{e(x)}$=0.92，$k$=4，代入(10—21)式可计算得：

$$\overline{S}_{\bar{y}} = \sqrt{\frac{0.8742}{12}\left[1+\frac{0.83}{0.92\times(4-1)}\right]} = 0.3078$$

SSR值与LSR值见表10—9。

表10—9　SSR值与LSR值表

秩次距k	2	3	4
$SSR_{0.05}$	2.86	3.01	3.10
$SSR_{0.01}$	3.82	3.99	4.10
$LSR_{0.05}$	0.883	0.929	0.957
$LSR_{0.01}$	1.179	1.232	1.266

各处理校正50日龄平均重多重比较结果见表10—10。

表10—10　各处理校正50日龄平均重多重比较表（SSR法）

处　理	$\overline{y}_{i.}^{'}$	$\overline{y}_{i.}^{'}$ -10.3514	$\overline{y}_{i.}^{'}$ -12.0758	$\overline{y}_{i.}^{'}$ -12.1821
配方3	12.3398	1.9884**	0.2640	0.1577
配方2	12.1821	1.8307**	0.1063	
配方1	12.0758	1.7244**		
对　照	10.3514			

多重比较结果表明：食欲添加剂配方3、2、1号的哺乳仔猪校正50 日龄平均重极显著高于对照，不同食欲添加剂配方间哺乳仔猪校正50日龄平均重差异不显著。

习　　题

1、何为试验控制？如何对试验进行统计控制？

2、什么是均积、协方差？均积与协方差有何关系？

3、对试验进行统计控制的协方差分析的步骤有哪些?

4、一饲养试验，设有两种中草药饲料添加剂和对照三处理，重复9次，共有27头猪参与试验，两个月增重资料如下。由于各个处理供试猪只初始体重差异较大，试对资料进行协方差分析。

中草药饲料添加剂对猪增重试验结果表　　（单位：*kg*）

处　理	2号添加剂		1号添加剂		对照组	
观测指标	初重*x*	增重*y*	初重*x*	增重*y*	初重*x*	增重*y*
观测值	30.5	35.5	27.5	29.5	28.5	26.5
	24.5	25.0	21.5	19.5	22.5	18.5
	23.0	21.5	20.0	18.5	32.0	28.5
	20.5	20.5	22.5	24.5	19.0	18.0
	21.0	25.5	24.5	27.5	16.5	16.0
	28.5	31.5	26.0	28.5	35.0	30.5
	22.5	22.5	18.5	19.0	22.5	20.5
	18.5	20.5	28.5	31.5	15.5	16.0
	21.5	24.5	20.5	18.5	17.0	16.0

（*b*=0.9832，线性回归关系极显著）。

5、四种配合饲料的比较试验，每种饲料各有供试猪 10 头，供试猪的初始重（kg）及试验后的日增重（kg）列于下表，试对试验结果进行协方差分析。

处　理	Ⅰ号料		Ⅱ号料		Ⅲ号料		Ⅳ号料	
观测指标	始重 *x*	增重 *y*	始重 *x*	增重 *y*	始重 *x*	增重 *y*	始重 *x*	增重 *y*
观测值	36	0.89	28	0.64	28	0.55	32	0.52
	30	0.80	27	0.81	22	0.62	27	0.58
	26	0.74	27	0.73	26	0.58	25	0.64
	23	0.80	24	0.67	22	0.58	23	0.62
	26	0.85	25	0.77	23	0.66	27	0.54
	30	0.68	23	0.67	20	0.55	28	0.54
	20	0.73	20	0.64	22	0.60	20	0.55
	19	0.68	18	0.65	23	0.71	24	0.44
	20	0.80	17	0.59	18	0.55	19	0.51
	16	0.58	20	0.57	17	0.48	17	0.51

（*b*=0.0073，线性回归关系极显著）

第十一章　非参数检验

前面有关章节讨论的参数检验都要求总体服从一定的分布，对总体参数的检验是建立在这种分布基础上的。例如，两样本平均数比较的 *t* 检验和多个样本平均数比较的 *F* 检验，都要求总体服从正态分布，推断两个或多个总体平均数是否相等。本章引入另一类检验——非参数检验（**non-parametric test**）。非参数检验是一种与总体分布状况无关的检验方法，它不依赖于总体分布的形式，应用时可以不考虑被研究的对象为何种分布以及分布是否已知。非参数检验主要是利用样本数据之间的大小比较及大小顺序，对两个或多个样本所属总体是否相同进行检验，而不对总体分布的参数如平均数、标准差等进行统计推断。当样本观测值的总体分布类型未知或知之甚少，无法肯定其性质，特别是观测值明显偏离正态分布，不具备参数检验的应用条件时，常用非参数检验。非参数检验具有计算简便、直观，易于掌握，检验速度较快等优点。

非参数检验法从实质上讲，只是检验总体分布的位置（中位数）是否相同，所以对于总体分布已知的样本也可以采用非参数检验法，但是由于它不能充分利用样本内所有的数量信息，检验的效率一般要低于参数检验方法。例如，非配对资料的秩和检验，其效率为 *t* 检验的 86.4%，就是说以相同概率判断出差异显著，*t* 检验所需的样本个数要少 13.6%。非参数检验内容很多，本章只介绍常用的符号检验（**sign test**），秩和检验（**rank-sum test**）和等级相关分析（**rank correlation analysis**）三种。

第一节　符号检验

一、配对资料的符号检验

（一）配对资料符号检验的意义　配对资料符号检验是根据样本各对数据之差的正负符号多少来检验两个总体分布位置的异同，而不去考虑差值的大小。每对数据之差为正值用“+”表示，负值用“－”表示。可以设想如果两个总体分布位置相同，则正或负出现的次数应该相等。若不完全相等，至少不应相差过大，否则超过一定的临界值就认为两个样本所来自的两个总体差异显著，分布的位置不同。显然这种检验比较的是中位数而不是平均数，当分布对称时，中位数与平均数相等。

（二）配对资料符号检验的基本步骤

1、提出无效假设与备择假设

H_O：甲、乙两处理差值 *d* 总体中位数=0；

H_A：甲、乙两处理差值 *d* 总体中位数≠0。

此时进行两尾检验。若将 H_A 中的“≠”改为“＜”或“＞”，则进行一尾检验。

2、计算差值并赋予符号　求甲、乙两个处理的配对数据的差值 *d*，*d*＞0 者记为“+”，*d*＜0 者记为“－”，*d*=0 记为“0”。统计“+”、“－”、“0”的个数，分别记为 n_+, n_-, n_0，令 $n = n_+ + n_-$。检验的统计量为 *K*，等于 n_+、n_- 中的较小者，即 $K = \min\{n_+, n_-\}$。

3、统计推断　由 n 查附表 11 符号检验用 K 临界值表（表中 $P_{(2)}$表示两尾概率，用于两尾检验，$P_{(1)}$表示一尾概率，用于一尾检验）得临界值 $K_{0.05(n)}$，$K_{0.01(n)}$。如果 $K>K_{0.05(n)}$，$P>0.05$，则不能否定 H_O，表明两个试验处理差异不显著；如果 $K_{0.01(n)}<K\leqslant K_{0.05(n)}$，$0.01<P\leqslant 0.05$，则否定 H_O，接受 H_A，表明两个试验处理差异显著；如果 $K\leqslant K_{0.01(n)}$，$P\leqslant 0.01$，则否定 H_O，接受 H_A，表明两个试验处理差异极显著（注意：当 K 恰好等于临界 K 值时，其确切概率常小于附表 11 中列出的相应概率）。

【例 11.1】某研究测定了噪声刺激前后 15 头猪的心率，结果见表 11-1。问噪声对猪的心率有无影响？

表 11-1　猪噪声刺激前后的心率（次/分钟）

猪　号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
刺激前	61	70	68	73	85	81	65	62	72	84	76	60	80	79	71
刺激后	75	79	85	77	84	87	88	76	74	81	85	78	88	80	84
差　值	-14	-9	-17	-4	1	-6	-23	-14	-2	3	-9	-18	-8	-1	-13
符　号	-	-	-	-	+	-	-	-	-	+	-	-	-	-	-

这是一个配对资料两尾检验的问题。

1、提出无效假设与备择假设

H_O ：噪声刺激前后猪的心率差值 d 总体中位数=0；

H_A ：噪声刺激前后猪的心率差值 d 总体中位数≠0。

2、计算差值并赋予符号　噪声刺激前后的差值及符号列于表 11-1 第 4 行和第 5 行，从而得 $n_+=2$、$n_-=13$，$n=n_++n_-=2+13=15$，$K=\min\{n_+,n_-\}=n_+=2$。

3、统计推断　当 n=15 时，查附表 11 得临界值 $K_{0.05(15)}$=3，$K_{0.01(15)}$=2，因为 K=2= $K_{0.01(15)}$，$P\leqslant 0.01$，表明噪声刺激对猪的心率影响极显著。

值得注意的是，虽然符号检验方法简单，但是，由于利用的信息较少，所以效率较低，且样本的配对数少于 6 时，不能检验出差别，在 7—12 时也不敏感，配对数在 20 以上时符号检验才较为有用。

二、样本中位数与总体中位数比较的符号检验

为了判断一个样本是否来自某已知中位数的总体，即样本所在总体的中位数是否等于某一已知总体的中位数，就需要进行样本中位数与总体中位数的差异显著性检验。其符号检验的基本步骤为：

1、提出无效假设与备择假设

H_O：样本所在的总体中位数=已知总体中位数；

H_A：样本所在的总体中位数≠已知总体中位数。

此时进行两尾检验。如果将备择假设 H_A 中的“≠”改为“＜”或“＞”，则进行一尾检验。

2、计算差值、确定符号及其个数　将样本各观测值中大于已知总体中位数者记为“+”，小于者记为“－”，等于者记为“0”。统计“+”、“－”、“0”的个数，分别记为 n_+、n_-、n_0，令 $n=n_++n_-$。假设检验的统计量 K 为 n_+、n_- 中的较小者，即 $K=\min\{n_+,n_-\}$。

3、统计推断　由 n 查附表 11 符号检验用 K 临界值表，得临界值 $K_{0.05(n)}$，$K_{0.01(n)}$。如

果 $K>K_{0.05(n)}$ ，P＞0.05，则不能否定 H_O，表明样本中位数与已知总体中位数差异不显著；如果 $K_{0.01(n)}<K\leqslant K_{0.05(n)}$ ，$0.01<P\leqslant 0.05$，则否定 H_O，接受 H_A，表明样本中位数与已知总体中位数差异显著；如果 $K\leqslant K_{0.01(n)}$，$P\leqslant 0.01$，则否定 H_O，接受 H_A，表明样本中位数与已知总体中位数差异极显著。

【例 11.2】已知某品种成年公黄牛胸围平均数为 140 厘米，今在某地随机抽取 10 头该品种成年公黄牛，测得一组胸围数字：128.1, 144.4, 150.3, 146.2, 140.6, 139.7, 134.1, 124.3, 147.9, 143.0（*cm*）。问该地成年公黄牛胸围与该品种胸围平均数是否有显著差异？

表 11-2 成年公黄牛胸围测定值符号检验表

牛号	1	2	3	4	5	6	7	8	9	10
胸围	128.1	144.4	150.3	146.2	140.6	139.7	134.1	124.3	147.9	143
差值	-11.9	4.4	6.3	6.2	0.6	-0.3	-5.9	-15.7	7.9	3
符号	-	+	+	+	+	-	-	-	+	+

1、提出无效假设与备择假设

H_O ：该地成年公黄牛胸围的平均数=140 厘米,

H_A ：该地成年公黄牛胸围的平均数≠140 厘米。

2、计算差值、确定符号及其个数　样本各观测值与总体平均数的差值及其符号列于表 11-3，并由此得 $n_+=6, n_-=4$，　$n=6+4=10$，　$K=\min\{n_+, n_-\}=n_-=4$。

3、统计推断　由 n=10，查附表 11，得 $K_{0.05(10)}$=1，$K>K_{0.05(10)}$ ，P＞0.05，不能否定 H_O，表明样本平均数与总体平均数差异不显著，可以认为该地成年公黄牛胸围的平均数与该品种胸围总体平均数相同。

第二节　　秩和检验

秩和检验也叫做符号秩和检验（**signed rank-sum tes**t），是一种经过改进的符号检验，或称 **Wilcoxon** 检验，其统计效率远较符号检验为高。因为它除了比较各对数据差值的符号外，还要比较各对数据差值大小的秩次高低。方法是通过将观测值按由小到大的次序排列，编定秩次，求出秩和进行假设检验。秩和检验与符号检验法不同，要求差数来自某些对称分布的总体，但并不要求每一差数来自相同的分布。

一、配对试验资料的符号秩和检验（Wilcoxon 配对法）

（一）基本步骤

1、提出无效假设与备择假设

H_O：差值 d 总体的中位数=0；

H_A：差值 d 总体的中位数≠0。

此时进行两尾检验。若将 H_A 中的“≠”改为“＜”或“＞”，则进行一尾检验。

2、编秩次、定符号　先求配对数据的差值 d，然后按 d 的绝对值从小到大编秩次。再根据原差值正负在各秩次前标上正负号，若差值 d=0，则舍去不记，若有若干个差值 d 的绝对值相等，则取其平均秩次。

3、确定统计量 T　分别计算正秩次及负秩次的和，并以绝对值较小的秩和绝对值为检验的统计量 T。

4、统计推断　记正、负差值的总个数为 n，根据 n 查附表 10(1)符号秩和检验用 T 临界值表，得 $T_{0.05(n)}$，$T_{0.01(n)}$。如果 $T>T_{0.05(n)}$，$P>0.05$，则不能否定 H_O，表明两个试验处理差异不显著；如果 $T_{0.01(n)}<T\leqslant T_{0.05(n)}$，$0.01<P\leqslant 0.05$，则否定 H_O，接受 H_A，表明两个试验处理差异显著；如果 $T\leqslant T_{0.01(n)}$，$P\leqslant 0.01$，则否定 H_O，接受 H_A，表明两个试验处理差异极显著（注意：当 T 恰好等于临界 T 值时，其确切概率常小于附表 10(1)中列出的相应概率）。

【例 11.3】某试验用大白鼠研究饲料维生素 E 缺乏与肝脏中维生素 A 含量的关系，先将大白鼠按性别、月龄、体重等配为 10 对，再把每对中的两只大白鼠随机分配到正常饲料组和维生素 E 缺乏饲料组，试验结束后测定大白鼠肝中维生素 A 的含量如表 11-4。试检验两组大白鼠肝中维生素 A 的含量是否有显著差异。

表 11-3 不同饲料鼠肝维生素 A 含量资料（国际单位/克）

鼠对别	1	2	3	4	5	6	7	8	9	10
正常饲料组	3550	2000	3100	3000	3950	3800	3620	3750	3450	3050
维生素E缺乏组	2450	2400	3100	1800	3200	3250	3620	2700	2700	1750
差值d_i	1100	-400	0	1200	750	550	0	1050	750	1300
秩次	+6	-1		+7	+3.5	+2		+5	+3.5	+8

1、提出无效假设与备择假设

H_O：差值 d 总体的中位数=0；

H_A：差值 d 总体的中位数≠0。

2、编秩次、定符号　计算表 11-3 中配对数据差值 d_i，将 d=0 的舍去，共有差值 n=8 个。按绝对值从小到大排列秩次并标上相应的符号，差值绝对值为 750 的有两个，它们的秩次为 3 和 4，所以其平均秩次为（3+4）/2=3.5，结果见表 11-3。

3、确定统计量 T　此例，正号有 7 个，其秩次为 2，3.5，3.5，5，6，7，8，秩次和为：2+3.5+3.5+5+6+7=35；

负号只有 1 个，其秩次为 1，秩次和等于 1。

负号秩次和较小，所以 T=1。

4、统计推断　由 n=8 查附表 10(1)得，$T_{0.05(8)}=3$，$T_{0.01(n)}=0$，因为 $T_{0.01(8)}<T<T_{0.05(8)}$，$0.01<P<0.05$，否定 H_O，接受 H_A，表明两个试验处理差异显著。

二、非配对试验资料的秩和检验（Wilcoxon 非配对法）

非配对试验资料的秩和检验是关于分别抽自两个总体的两个独立样本之间秩和的成组比较，它比配对资料的秩和检验的应用更为普遍。

（一）基本步骤

1、提出无效假设与备择假设

H_O：甲样本所在的总体的中位数=乙样本所在的总体的中位数；

H_A：甲样本所在的总体的中位数≠乙样本所在的总体的中位数。

此时进行两尾检验。若将 H_A 中的“≠”改为“＜”或“＞”，则进行一尾检验。

2、求两个样本合并数据的秩次　假设两个样本的含量分别为 n_1 和 n_2，则将两样本的观测值合并后，总的数据为 n_1+n_2 个。将合并后的数据按从小到大的顺序排列，与每个数据对应的序号即为该数据的秩次，最小数值的秩次为“1”，最大数值的秩次为“n_1+n_2”。遇不同样本的相同观测值时，其秩次取原秩次的平均秩次，但是同一样本内遇相同的观测值时则不必求平均秩次，秩次孰先孰后都可以。

3、确定统计量 T　将两个样本重新分开，并计算各自的秩和。将较小的那个样本含量作为 n_1，其秩和作为检验的统计量 T。若 $n_1=n_2$，则任取一组的秩和为 T。

4、统计推断　由 n_1、(n_2 －n_1)查附表 10(3)，得接受区域$T'_{0.05}-T_{0.05}$，$T'_{0.01}-T_{0.01}$。若 T 在$T'_{0.05}-T_{0.05}$之内，$P>0.05$，则不能否定 H_O，表明两个试验处理差异不显著；若 T 在$T'_{0.05}-T_{0.05}$之外但在$T'_{0.01}-T_{0.01}$之内，$0.01<P\leqslant 0.05$，则否定 H_O，接受 H_A，表明两个试验处理差异显著；若 T 在$T'_{0.01}-T_{0.01}$之外，$P<0.01$，则否定 H_O，接受 H_A，表明两个试验处理差异极显著。

【例 11.4】　研究两种不同能量水平饲料对 5-6 周龄肉仔鸡增重（克）的影响，资料如表 11-4 所示。问两种不同能量水平的饲料对肉仔鸡增重的影响有无差异？

表 11-4 两种不同能量水平饲料的肉仔鸡增重及秩和检验

饲　料	肉仔鸡增重（g）									
高能量	603	585	598	620	617	650				n_1=6
秩　次	12	8.5	11	14	13	15				T_1=73.5
低能量	489	457	512	567	512	585	591	531	467	n_2=9
秩　次	3	1	4	7	5	8.5	10	6	2	T_2=46.5

1、提出无效假设与备择假设

H_O：高能量饲料增重总体的中位数=低能量饲料增重总体的中位数；

H_A：高能量饲料增重总体的中位数≠低能量饲料增重总体的中位数。

2、编秩次　将两组数据混合从小到大排列为秩次。在低能量组有两个“512”，不求平均秩次，其秩次分别为 4 和 5；在高、低两组有一对数据为“585”，需求它们的平均秩次：(8+9)/2=8.5。结果见表 11-4。

3、确定统计量 T　以较小样本的秩次和为统计量 T，即 T= 73.5。

4、统计推断　由 n_1=6, n_2–n_1=9–6=3 查附表 10(3)得，$T'_{0.05}-T_{0.05}$ 为 31—65，$T'_{0.01}-T_{0.01}$ 为 26—70。T=73.5 在$T'_{0.01}-T_{0.01}$，即 26—70 之外，$P<0.01$，否定 H_O，接受 H_A，表明饲料能量高低对肉仔鸡增重的影响差异极显著。

三、多个样本比较的秩和检验（Kruskal-Wallis 法，H 法）

多个样本比较的秩和检验的 **Kruskal-Wallis** 法，又称 H 检验法。该法的前提是假设抽样总体是连续的和相同的，利用多个样本的秩和来推断它们分别代表的总体之分布位置是否相同，检验的基本步骤是：

1、提出无效假设与备择假设

H_O：各个样本所分别代表的各总体分布位置相同；

H_A：各个样本所分别代表的各总体分布位置不完全相同。

2、编秩次、求秩和　将各个样本的所有观测值混合后，按照由小到大的顺序排成 1，2，…，n 个秩次。不同样本的相同观测值，取平均秩次；一个样本内的相同观测值，不求平均秩次。按样本把每个观测值的秩次一一相加，求出各样本的秩和。

3、求 H 值

$$H=\frac{12}{n(n+1)}\sum\frac{R_i^2}{n_i}-3(n+1) \qquad (11\text{-}1)$$

式中，R_i 为第 i 个样本的秩次之和；

n_i 为第 i 个样本的含量；$n=\Sigma n_i$

4、统计推断　根据 n, n_i 查附表 10（2），得临界值：$H_{0.05}$，$H_{0.01}$。若 $H<H_{0.05}$，$P>0.05$，不能否定 H_O，可以认为各样本代表的各总体分布位置相同；若 $H_{0.05}\leqslant H<H_{0.01}$，$0.01<P\leqslant 0.05$，否定 H_O，接受 H_A，表明各样本所代表的各总体分布位置显著不同；若 $H\geqslant H_{0.01}$，$P\leqslant 0.01$，表明各样本所代表的各总体分布位置极显著不同。

当样本数 $k>3$，$n_i>5$ 时，不能从附表 10（2）中查得 H 值。这时 H 近似地呈自由度为 $k-1$ 的 χ^2 分布，可对 H 进行 χ^2 检验。

当相同的秩次较多时，按（11-1）式计算的 H 值常常偏低，此时应按（11-2）式求校正的 H 值 H_C：

$$H_C=\frac{H}{\left[1-\frac{\sum(t_j^3-t_j)}{n^3-n}\right]} \qquad (11\text{-}2)$$

式中，t_j 表示某个数重复的次数。

【例 11.5】某试验研究三种不同制剂治疗钩虫的效果，用 11 只大白鼠做试验，分为三组。每只鼠先人工感染 500 条钩蚴，感染后第 8 天，三组分别给服用甲、乙、丙三种制剂，第 10 天全部解剖检查各鼠体内活虫数，试验结果如表 11-5 所示。试检验三种制剂杀灭钩虫的效果有无差异。

表 11-7　三种制剂杀灭钩虫效果及秩和检验

制剂甲组（a）		制剂乙组（b）		制剂丙组（c）	
活虫数	秩次	活虫数	秩次	活虫数	秩次
279	6	229	4	210	3
338	11	274	5	285	7
334	10	310	9	117	1
198	2				
303	8				
n_i	5		3		3
R_i	37		18		11

1、提出无效假设与备择假设

H_O：三种制剂活虫数总体分布位置相同；

H_A：三种制剂活虫数总体分布位置不完全相同。

2、编秩次、求秩和　三个组观测值混合后的秩次如表 11-5 所示，最后一行为各组

秩次之和。

3、求 H 值　由（11-1）式，得

$$H=\left[\frac{12}{11(11+1)}\left(\frac{37^2}{5}+\frac{18^2}{3}+\frac{11^2}{3}\right)\right]-3(11+1)=2.38$$

4、统计推断　当 n=11, n_1=5, n_2=3, n_3=3 时，查附表 10（2），得 $H_{0.05}$=5.65。因为 $H<H_{0.05}, P>0.05$，不能否定 H_O，表明三种制剂杀灭钩虫的效果差异不显著。

【例 11.6】对某种疾病采用一穴、二穴、三穴作针刺治疗，治疗效果分为控制、显效、有效、无效 4 级。治疗结果见表 11-6 第（2）、（3）、（4）栏。问 3 种针刺治疗方式疗效有无显著差异？

表 11-6　3 种针刺方式治疗效果及秩和检验

等　级	一穴	二穴	三穴	合计	秩次范围	平均秩次	各组秩和		
							一穴	二穴	三穴
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
控　制	21	30	10	61	1~61	31.0	651.0	930.0	310.0
显　效	18	10	22	50	62~111	86.5	1557.0	865.0	1903.0
有　效	15	8	11	34	112~145	128.5	1927.5	1028.0	1413.5
无　效	5	2	8	15	146~160	153.0	765.0	306.0	1224.0
合　计	59	50	51	160			4900.5	3129.0	4850.5
	（n_1）	（n_2）	（n_3）	（n）			（R_1）	（R_2）	（R_3）

1、提出无效假设与备择假设

H_O：三种针刺方式治疗效果相同；

H_A：三种针刺方式治疗效果不完全相同。

2、编秩次、求秩和　秩次、秩和等的计算结果列于表 11-6。其中的合计栏（5）=（2）+（3）+（4）栏；秩次范围栏（6）为每一等级组应占的秩次；平均秩次栏（7），是因为同一组所包含的秩次同属一个等级，不能分列出高低，故一律以其平均秩次为代表，平均秩次等于各等级组秩次下限与上限之和的平均；各组秩和 R_1、R_2、R_3 分别等于第（2）、（3）、（4）栏乘以第（7）栏所得第（8）、（9）、（10）栏各自的和。

3、求 H 值　因为各等级组段均以平均秩次作为代表，视为相同秩次，其相同秩次的个数 t_j 等于各自的秩次合计，见第（5）栏。显然相同秩次较多，宜用（11-2）式求 H_C。先按（11-1）式计算 H 值：

$$H=\frac{12}{160\times(160+1)}\left(\frac{4900.5^2}{59}+\frac{3129.0^2}{50}+\frac{4850.5^2}{51}\right)-3\times(160+1)=12.7293$$

而　$$\sum\left(t_j^3-t_j\right)=(61^3-61)+(50^3-50)+(34^3-34)+(15^3-15)=394500$$

于是利用（11-2）式，得：

$$H_C=\frac{12.7293}{1-\frac{394500}{160^3-160}}=\frac{12.7293}{0.9037}=14.0858$$

此试验处理数为 3，所以 df=3－1=2，查 χ^2 值表得 $\chi^2_{0.01(2)}=9.21$。因为 $H_C>\chi^2_{0.01(2)}$，

$P<0.01$,表明 3 种针刺方式的治疗效果差异极显著。

四、多个样本两两比较的秩和检验（Nemenyi-Wilcoxson-Wilcox 法）

当多组计量资料或等级资料经多个样本比较的秩和检验，认为各总体的分布位置不完全相同时，常需要进一步作两两比较的秩和检验，以推断哪两个总体的分布位置不同，哪两个总体分布位置并无不同。这个方法类似方差分析中的多重比较，常用 q 法：

$$q=\frac{R_i-R_j}{S_{R_i-R_j}} \tag{11-3}$$

式中，$S_{R_i-R_j}$ 为秩和差异标准误，计算公式为：

$$S_{R_i-R_j}=\sqrt{\frac{n(nk)(nk+1)}{12}} \tag{11-4}$$

n 为样本含量即处理的重复数；k 为比较的两秩和差数范围内所包含的处理数。可见，这里的 q 法只适用于重复数相等的试验资料。

计算 q 值后，以 $df=\infty$ 和 k 查附表 5，得临界值 $q_{\alpha(\infty,k)}$，作出统计推断。

【例 11.7】 某种激素 4 种剂量对大白鼠耻骨间隙宽度增加量的影响试验，结果见表 11-7。问 4 种剂量大白鼠耻骨间隙的增加量是否有显著差异？

表 11-7　四种剂量大白鼠耻骨间隙增加量及秩和检验

剂 量	增加量（单位：mm）										R_i
1	0.15	(1)	0.30	(2)	0.40	(3)	0.40	(4)	0.50	(5)	15
2	1.20	(6.5)	1.35	(8)	1.40	(9.5)	1.50	(11)	1.90	(14)	49
3	2.50	(19.5)	1.20	(6.5)	1.40	(9.5)	2.00	(15)	2.20	(16.5)	67
4	1.80	(13)	1.60	(12)	2.50	(19.5)	2.20	(16.5)	2.30	(18)	79

1、提出无效假设与备择假设

H_O：四种剂量大白鼠耻骨间隙宽度增加量的总体分布位置相同；

H_A：四种剂量大白鼠耻骨间隙宽度增加量的总体分布位置不全相同。

2、编秩次、求秩和　将四组观测值混合，由小到大编秩次，见表 11-7 括号内数字。不同组的相同观测值取平均秩次，如第 2、3 组各有一个 1.20，取它们原来秩次 6 和 7 的平均 6.5，余此类推；同一组内相同观测值不求平均秩次。各组秩和见表 11-7 最后一栏。

3、求 H 值　因为本例有 2 个 1.20，2 个 1.40，2 个 2.20，2 个 2.50 ，所以用（11-2）式求校正 H_C。先按（11-2）式计算 H。

$$H=\frac{12}{20(20+1)}\left(\frac{15^2}{5}+\frac{49^2}{5}+\frac{67^2}{5}+\frac{79^2}{5}\right)-3(20+1)=13.32$$

而

$$\sum(t_j^3-t_j)=(2^3-2)+(2^3-2)+(2^3-2)+(2^3-2)=24$$

所以

$$H_C=\frac{13.32}{1-\frac{24}{20^3-20}}=\frac{13.32}{0.9970}=13.36$$

4、统计推断　本例 k=4，超出附表 10（2）的范围，故用 χ^2 值（附表 7）进行统计推断。当 df=4－1=3 时，查附表 7，得 $\chi^2_{0.01(3)}=11.34$。因为 $H_C>\chi^2_{0.01(3)}$，P＜0.01，表明用 4 种剂量的大白鼠耻骨间隙宽度的增加量差异极显著。

5、多个样本的两两比较　列出两两比较表（表 11-8）。

表 11-8　4 种剂量大白鼠耻骨间隙宽度增加量秩和两两比较

比　较	差数 R_i-R_j	秩次距 k	$S_{R_i-R_j}$	q 值	临界 q 值		检验结果
					α=0.05	α=0.01	
（1）	（2）	（3）	（4）	（5）	（6）	（7）	（8）
1 与 4	64	4	13.2288	4.84	3.63	4.40	**
1 与 3	52	3	10.0000	5.20	3.32	4.12	**
1 与 2	34	2	6.7700	5.02	2.77	3.64	**
2 与 4	20	3	10.0000	2.00	3.32	4.12	*ns*
2 与 3	18	2	6.7700	2.66	2.77	3.64	*ns*
3 与 4	12	2	6.7700	1.77	2.77	3.64	*ns*

用表 11-7 中相应的秩和 R 栏求秩和差数 R_i-R_j，见第（2）栏；确定秩次距 k，例如 1 与 4 的比较，其秩和差数 64 范围内有 4 个处理，k=4；1 与 3 的比较，其秩和差数 52 范围内有 3 个处理，k=3，余此类推，见第（3）栏。利用（11-4）式计算各秩和差异标准误 $S_{R_i-R_j}$：

k=4 时，　$S_{R_1-R_4}=\sqrt{\frac{5(5\times4)(5\times4+1)}{12}}=13.2288$

k=3 时，　$S_{R_1-R_3}=\sqrt{\frac{5(5\times3)(5\times3+1)}{12}}=10.0000$

k=2 时，　$S_{R_1-R_3}=\sqrt{\frac{5(5\times2)(5\times2+1)}{12}}=6.7700$

见第（4）栏。用（11-3）式计算 q 值，即(2)/(4)得（5）栏。然后根据 df=∞和 k 查临界 q 值，列于第（6）、（7）栏。当 q＜1 者，差异必不显著，该行的临界 q 值不必列出。最后将各 q 值与相应的临界 q 值比较，作出统计推断。检验结果表明：第 1 种剂量与第 2、3、4 种剂量差异极显著；第 2 种剂量与第 3、4 种剂量，第 3 种剂量与第 4 种剂量差异不显著。

第三节　　等级相关分析

第八章所述的相关、回归分析法适用于变量为正态分布的资料。在实际工作中，经常遇到有些资料并不呈正态分布。对于这样的资料的分析只能用非参数法。分析两个变量间是否相关的非参数法，最常用的是等级相关分析。

等级相关是一种分析 x、y 两个变量的等级间是否相关的方法。先按 x、y 两变量的大小次序，分别由小到大编上等级（秩次），再看两个变量的等级间是否相关。等级相关程度的大小和相关性质用等级相关系数（**coefficient of rank correlation**）表示。等级相关系数亦称为秩相关系数。样本等级相关系数记为 r_s，它是总体等级相关系数 ρ_S 的估计值。等级相关系数 r_s 具有与相关系数 r 相同的特性，它的值介于-1 与 1 之间，r_s 为正表示正相关，r_s 为负表示负相关，r_s 等于零为零相关。常用的等级相关分析方法有 **Spearman** 等级相关和 **Kendall** 等级相关等，本节只介绍 **Spearman** 等级相关系数的计算及其显著性检验。其基本分析步骤是：

1、 计算等级相关系数 r_s 先将变量 x、y 分别由小到大列出等级，相邻两数相同时，取平均等级；再求出每对等级之差 d，利用（11-5）式计算等级相关系数：

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} \qquad (11\text{-}5)$$

式中，n 为变量的对子数，d 秩次之差。

当相同秩次较多时，会影响 $\sum d^2$ 值，应采用（11-6）式计算校正的等级相关系数 r_s'：

$$r_s' = \frac{\frac{n^3-3}{6} - (t_x + t_y) - \sum d^2}{\sqrt{\left(\frac{n^3-n}{6} - 2t_x\right)\left(\frac{n^3-n}{6} - 2t_y\right)}} \qquad (11\text{-}6)$$

式中，t_x、t_y 的计算公式相同，均为：$\sum \frac{t_i^3 - t_i}{12}$。在计算 t_x 时，t_i 为 x 变量的相同秩次数；在计算 t_y 时，t_i 为 y 变量的相同秩次数。

2、r_s 的显著性检验

（1）提出无效假设与备择假设 H_O：$\rho_S = 0$；H_A：$\rho_S \neq 0$

（2）统计推断 根据 n 查附表 12，得临界 $r_{s(\alpha)}$ 值。若 $|r_s| < r_{s(0.05)}$，$P > 0.05$，不能否定 H_O，表明两变量 x、y 等级相关不显著；若 $r_{s(0.05)} \leqslant |r_s| < r_{s(0.01)}$，$0.01 < P \leqslant 0.05$，否定 H_O，接受 H_A，表明两变量 x、y 等级相关显著；若 $|r_s| \geqslant r_{s(0.01)}$，$P \leqslant 0.01$，否定 H_O，接受 H_A，表明两变量 x、y 等级相关极显著。

【例 11.8】研究含有必需氨基酸添加剂的某种饲料的营养价值时，用大白鼠做试验获得了关于进食量（x）和增重（y）的数据，见表 11-9。试分析大白鼠的进食量与增重之间有无相关。

表 11-9　大白鼠进食量与增重结果及等级相关分析表

鼠　号	变量 x		变量 y		秩次差	秩次差平方
	进食量（g）	秩次	增重（g）	秩次	d	d^2

1	820	7.5	165	7	0.5	0.25
2	780	5	158	5.5	-0.5	0.25
3	720	4	130	2	2	4
4	867	9	180	9	0	0
5	690	3	134	3	0	0
6	787	6	167	8	-2	4
7	934	10	186	10	0	0
8	679	2	145	4	-2	4
9	639	1	120	1	0	0
10	820	7.5	158	5.5	2	4
合　计						16.5

1、计算等级相关系数 r_s　对表 11-9 中各个试验数据分别按进食量与增重从小到大，排列秩次，对数值相同的数据则取平均秩次，如进食量 820 克的平均秩次为（7+8）/2=7.5。求出进食量的秩次与增重的秩次之差 d 和秩次差平方 d^2。利用（11-5）式，得

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 16.5}{10(10^2-1)} = 0.90$$

2、r_s 的显著性检验　此例 n=10，查附表 12，得 $r_{s(0.01)}$=0.794, 因为 $r_s > r_{s(0.01)}$，$P<0.01$，等级相关极显著，表明大白鼠的进食量与增重之间存在着极显著正相关。

【例 11.9】某研究观察雌鼠的年龄（月）与所产仔鼠的初生重（克）之间的关系，得表 11-10 的结果，问仔鼠初生重与雌鼠年龄之间是否存在相关？

表 11-10　雌鼠年龄和仔鼠初生重的等级相关分析表

编号 (1)	雌鼠年龄(月) (2)	仔鼠初生重(g) (3)	年龄秩次 (4)	初生重秩次 (5)	秩次差(d) (6)	秩次差平方(d^2) (7)
1	12	19	10	10	0	0
2	7	13	5.5	7.5	-2	4
3	4	8	2.5	1.5	1	1
4	9	8	8.5	1.5	7	49
5	7	13	5.5	7.5	-2	4
6	2	14	1	9	-8	64
7	9	12	8.5	5.5	3	9
8	5	10	4	3	1	1
9	8	12	7	5.5	1.5	2.25
10	4	11	2.5	4	-1.5	2.25
合计						136.5

1、计算等级相关系数 r_s　先确定或计算雌鼠年龄和仔鼠出生重两个变量的秩次、秩次差、秩次差的平方，结果见表 11-10 第（4）、（5）、（6）、（7）栏。

本例年龄秩次有 2 个 5.5，2 个 2.5，2 个 8.5；初生重秩次有 2 个 7.5，2 个 1.5，2 个 5.5，相同的秩次较多，用校正公式（11-6）式计算等级相关系数。

先计算 t_x 和 t_y，对于 t_x，t_1=2，t_2=2，t_3=2；对于 t_y，t_1=2，t_2=2，t_3=2：

$$t_x = \sum \frac{t_i^3 - t_i}{12} = \frac{(2^3-2)+(2^3-2)+(2^3-2)}{12} = 1.5$$

$$t_y=\sum\frac{t_i^3-t_i}{12}=\frac{(2^3-2)+(2^3-2)+(2^3-2)}{12}=1.5$$

于是：

$$r_s^{'}=\frac{\frac{10^3-3}{6}-(1.5+1.5)-136.5}{\sqrt{\left(\frac{10^3-10}{6}-2\times1.5\right)\left(\frac{10^3-10}{6}-2\times1.5\right)}}=\frac{26.6667}{\sqrt{(165-3)(165-3)}}=0.1646$$

2、$r_s^{'}$的显著性检验　此例 n=10，查附表 12，得$r_{s(0.05)}$=0.648，$r_s^{'}<r_{s(0.05)}$，P＞0.05，表明雌鼠的年龄与仔鼠初生重相关不显著。

习　题

1、参数检验与非参数检验有何区别？各有什么优缺点？

2、为什么在秩和检验编秩次时不同组间出现相同数据要给予“平均秩次”，而同一组的相同数据不必计算“平均秩次”？

3、两样本比较的秩和检验的检验假设是否可用 $\mu_1=\mu_2$ 表示？为什么？

4、今测定了 10 头猪进食前后血糖含量变化如下表，分别用配对资料的符号检验和秩和检验法检验进食后血糖的平均含量差异是否显著？

猪号	1	2	3	4	5	6	7	8	9	10
饲前	120	110	100	130	123	127	118	130	122	145
饲后	125	125	120	131	123	129	120	129	123	140

（符号检验 P＞0.05，配对秩和检验 T=8.5，P＞0.05）

5、已知某地正常人尿氟含量的中位数为 0.86*mg/L*。今在该地某厂随机抽取 12 名工人，测得尿氟含量如下：0.84, 0.86, 0.88, 0.94, 0.97, 1.01, 1.05, 1.09, 1.20, 1.35, 1.83（mg/L）。问该厂工人的尿氟含量是否显著高于当地正常人？（K=1, P≤0.05）

6、将一种生物培养物以等量分别接种到两种综合培养基 *A* 和 *B* 上，共接种 10 瓶 *A* 培养基和 15 瓶 *B* 培养基。一周后计算培养壁上单位面积的生物培养物细胞平均贴壁数，获得试验数据如下：

培养基 *A*	254	140	193	153	316	473	389	257	167	147
培养基 *B*	331	257	478	339	407	396	144	357	287	568
	483	396	245	403	390					

试检验两种培养基的培养效果有无显著差异？（T=84.5, P＞0.05）

7、将未达到性成熟的雌性大家鼠 14 只，随机分为三组，分别为 5 只、5 只和 4 只，各组分别注射剂量为 0.64μg/鼠 1.64μg/鼠和 2.64μg/鼠的促性腺激素，每天一次，连续注射三天后将其杀死，取出卵巢称重。试验结果见下表：

处　　理 （注射剂量μg/鼠）	I （0.64μg /鼠）	II （1.64μg/鼠）	III （2.64μg/鼠）
卵巢重量 (*mg*)	16.5	24.9	41.8
	45.0	51.6	54.6
	26.5	35.7	31.5
	32.9	33.6	39.9
	20.0	30.4	

问三种不同剂量促性腺激素对大家鼠卵巢增重效果是否有差异？（H=3.21，P＞0.05）

8、观察三个品种母猪乳头数得如下次数分布表。问三个品种母猪乳头数有无差异？

乳头数（个）	母猪数（头）			合　计
	品种 *A*	品种 *B*	品种 *C*	
＜12	1	2	1	4
13	3	2	2	7
14	2	5	4	11
15	4	3	2	9
16	4	6	3	13
17	3	4	2	9
＞18	2	1	2	5

（H_C=0.102, P＞0.05）

9、四种抗菌素的抑菌效力比较研究，以细菌培养皿内抑菌区直径为指标，并获得如下结果：

平皿号	抗菌素Ⅰ	抗菌素Ⅱ	抗菌素Ⅲ	抗菌素Ⅳ
1	28	23	24	19
2	27	25	20	22
3	29	24	22	21
4	26	24	21	23
5	28	23	23	22

试检验四种抗菌素的抑菌效力有无显著差异？如果有显著差异，作两两比较。

（H_C=14.534，P＜0.01，两两比较只有第Ⅲ与第Ⅳ组差异不显著＝

10、用最佳线性无偏预测（*BLUP*）法和相对育种值（*RBV*）法对 12 头肉牛种公牛的种用价值作评定，其评定结果排序如下。问两种评定方法是否显著相关？

序　号	1	2	3	4	5	6	7	8	9	10	11	12
BLUP 法	9 号	8 号	5 号	4 号	10 号	11 号	3 号	6 号	12 号	2 号	1 号	7 号
RBV 法	9 号	8 号	4 号	5 号	10 号	11 号	6 号	3 号	12 号	2 号	1 号	7 号

（$r_s = 0.9860$，P＜0.01）

11、有甲乙二鉴定员，对 7 头贫乏饲养 3 周的大白鼠评定的等级如下表。问甲、乙两人评定结果是否相似？

序 号	1	2	3	4	5	6	7
甲	4 号	1 号	6 号	5 号	3 号	2 号	7 号
乙	4 号	2 号	5 号	6 号	1 号	3 号	7 号

（$r_s = 0.8571$，　P＜0.01）

第十二章　试验设计

试验设计（**experimental design**）是数理统计学的一个分支，是进行科学研究的重要工具。由于它与生产实践和科学研究紧密结合，在理论和方法上不断地丰富和发展，因而广泛地应用于各个领域。

第一节　试验设计概述

一、试验设计的基本概念

试验设计，广义理解是指试验研究课题设计，也就是整个试验计划的拟定。主要包括课题的名称、试验目的，研究依据、内容及预期达到的效果，试验方案，试验单位的选取、重复数的确定、试验单位的分组，试验的记录项目和要求，试验结果的分析方法，经济效益或社会效益估计，已具备的条件，需要购置的仪器设备，参加研究人员的分工，试验时间、地点、进度安排和经费预算，成果鉴定，学术论文撰写等内容。而狭义的理解是指试验单位（如动物试验的畜、禽）的选取、重复数目的确定及试验单位的分组。生物统计中的试验设计主要指狭义的试验设计。

试验设计的目的是避免系统误差，控制、降低试验误差，无偏估计处理效应，从而对样本所在总体作出可靠、正确的推断。

试验设计的任务是在研究工作进行之前，根据研究项目的需要，应用数理统计原理，作出周密安排，力求用较少的人力、物力和时间，最大限度地获得丰富而可靠的资料，通过分析得出正确的结论，明确回答研究项目所提出的问题。如果设计不合理，不仅达不到试验的目的，甚至导致整个试验的失败。因此，能否合理地进行试验设计，关系到科研工作的成败。

二、动物试验的任务

在畜牧、水产等试验研究中，通常以动物作为试验对象，因而将所进行的试验统称为动物试验。它的主要任务在于研究、揭示和掌握动物生长发育规律、及这些规律与饲养管理、环境条件等的关系。通过试验，鉴定新的动物品种（系），探索新的饲料配方，饲养管理方法和技术措施，找出其中的规律，并将这些规律应用到生产实践中去，以解决畜牧业、水产业等生产中存在的问题，进一步提高产品的质量和数量，取得更大的经济效益和社会效益，从而推动畜牧业、水产业等事业的发展。

三、动物试验的特点与要求

在动物试验研究中，除小部分可在严格控制的试验条件下进行外，大部分试验都与外界环境接触或要在外界环境中进行，试验的对象是生长在不同时期、各种环境中的动物。因此，

动物试验结果除有试验处理的作用外，还要受到许多其它因素的干扰和制约，这些因素对试验结果可以产生较大的影响。所以，我们要在充分认识这些干扰因素的情况下，对其进行合理、有效的控制，以保证试验结果的正确性。

（一）动物试验的特点

1、试验干扰因素多　首先是动物本身存在差异，这种差异是试验中误差的重要来源。例如，在同一饲养试验中，为使供试动物均匀一致，要选择到遗传来源一致、同年龄、同体重、同性别的动物进行试验是比较困难的；其次，自然环境如温度、湿度、光照、通风等存在差异，不能完全控制一致；第三，饲养管理条件存在差异，如在试验过程中的管理方法、饲养技术、畜舍笼位的安排等不一致；第四，试验人员操作技术上的差异，如对试验动物的性状、指标进行测量时，时间、人员和仪器等不完全一致。

2、试验具有复杂性　在畜牧、水产等动物试验中所研究的各种试验对象，它们都有自己的生长发育规律和遗传特性，并与环境、饲养管理等条件密切相关，而且这些因素之间又相互影响，相互制约，共同作用于供试对象。所以在试验中，人们不可能做到对环境条件等一一加以控制，当然也就不易精确地分析出各个因素的单独作用。因此，在多变的各种条件下，不能只依据少数的或短期的试验，而必须经过不同条件下的一系列试验，才能获得比较正确的结果。

3、试验周期长　动物完成一个生活世代的时间较长，特别是大动物、单胎动物、具有明显季节性繁殖的动物更为突出。因此，有的一年内不能进行多次试验，例如动物遗传育种试验，有的需用几年的时间才能完成整个试验。应尽量克服周期长、试验年度间差异的影响，以获得正确的结论。

（二）动物试验的基本要求　由于动物试验具有上述特点，为了保证试验的质量，在试验中应尽可能地控制和排除非试验因素的干扰，合理地进行试验设计、准确地进行试验，从而提高试验的可靠程度，使试验结果在生产实际中真正发挥作用。为此，对动物试验有以下几点要求：

1、试验要有代表性　动物试验的代表性包括生物学和环境条件两个方面的代表性。生物学的代表性，是指作为主要研究对象的动物品种、个体的代表性，并要有足够的数量。例如，进行品种的比较试验时，所选择的个体必须能够代表该品种，不要选择性状特殊的个体，并根据个体均匀程度，在保证试验结果具有一定可靠性的条件下，确定适当的动物数量。环境条件的代表性是指代表将来计划推广此项试验结果的地区的自然条件和生产条件，如气候、饲料、饲养管理水平及设备等。代表性决定了试验结果的可利用性，如果一个试验没有充分的代表性，再好的试验结果也不能推广和应用，就失去了实用价值。

2、试验要有正确性　试验的正确性包括试验的准确性和试验的精确性。在进行试验的过程中，应严格执行各项试验要求，将非试验因素的干扰控制在最低水平，以避免系统误差，降低试验误差，提高试验的正确性。

3、试验要有重演性　重演性是指在相同条件下，重复进行同一试验，能够获得与原试验相类似的结果，即试验结果必须经受得起再试验的检验。试验的目的在于能在生产实践中推广试验结果，如果一个在试验中表现好的结果在实际生产中却表现不出来，那么，试验就失去了意义。由于试验受供试动物个体之间差异和复杂的环境条件等因素影响，不同地区或不同时间进行的相同试验，结果往往不同；即使在相同条件下的试验，结果也有一定出入。因此，为了保证试验结果的重演性，必须认真选择供试动物，严格把握试验过程中的各个环节，在有条件的情况下，进行多年或多点试验，这样所获得的试验结果才具有较好的重演性。

第二节　动物试验计划

一、试验计划的内容及要求

进行任何一项科学试验，在试验前必须制定一个科学的、全面的试验计划，以便使该项研究工作能够顺利开展，从而保证试验任务的完成。虽然科研项目的级别、种类等有所不同，但基本要求是一致的。试验计划的内容一般应包括以下几个方面：

（一）课题名称与试验目的　科研课题的选择是整个研究工作的第一步。课题选择正确，此项研究工作就有了很好的开端。一般来说，试验课题通常来自两个方面。一是国家或企业指定的试验课题，这些试验课题不仅确定了科研选题的方向，而且也为研究人员选题提供了依据，并以此为基础提出最终的目标和题目。二是研究人员自己选定的试验课题。研究人员自选课题时，首先应该明确为什么要进行这项科学研究，也就是说，应明确研究的目的是什么，解决什么问题，以及在科研和生产中的作用、效果如何等。例如，畜禽口服补液盐对雏鸡的影响试验，主要目的在于提高雏鸡的成活率。若用于肉鸡，则目的在于促进增重。

选题时应注意以下几点：

1、实用性　要着眼于畜牧、水产等科研和生产中急需解决的问题，同时从发展的观点出发，适当照顾到长远或不久将来可能出现的问题。

2、先进性　在了解国内外该研究领域的进展、水平等基础上，选择前人未解决或未完全解决的问题，以求在理论、观点及方法等方面有所突破。

3、创新性　研究课题要有自己的新颖之处。

4、可行性　就是完成科研课题的可能性，无论是从主观条件方面，还是客观条件方面，都要能保证研究课题的顺利进行。

（二）研究依据、内容及预期达到的经济技术指标　课题确定后，通过查阅国内外有关文献资料，阐明项目的研究意义和应用前景，国内外在该领域的研究概况、水平和发展趋势，理论依据、特色与创新之处。详细说明项目的具体研究内容和重点解决的问题，以及取得成果后的应用推广计划，预期达到的经济技术指标及预期的技术水平等。

（三）试验方案和试验设计方法　试验方案是全部试验工作的核心部分，主要包括研究的因素、水平的确定等，具体内容详述于后。方案确定后，结合试验条件选择合适的试验设计方法，具体内容详见本章第四节至第八节。

（四）供试动物的数量及要求　试验动物或试验对象选择正确与否，直接关系到试验结果的正确性。因此，试验动物应力求比较均匀一致，尽量避免不同品种、不同年龄、不同胎次、不同性别等差异对试验的影响。新引进的动物应有一个适应和习惯过程。试验动物的数量，可根据本章第十节介绍的方法来计算。

（五）试验记录的项目与要求　为了收集分析结果需要的各个方面资料，应事先以表格的形式列出需观测的指标与要求，例如，饲养试验中的定期称重，定期为 1 周、10 天或半月称重，称重一般在清晨空腹或喂前进行等。

（六）试验结果分析与效益估算　试验结束后，对各阶段取得的资料要进行整理与分析，所以应明确采用统计分析的方法，如 t 检验，方差分析、回归与相关分析等。每一

种试验设计都有相应的统计分析方法，统计方法应用不恰当，就不能获得正确的结论。如果试验效果显著，同时应计算经济效益。如某农场为饲养肉用仔鸡而配制的“维生素添加剂”的试验，不仅记录分析它对生长发育的效果，而且还计算出喂青料（对照组）每只鸡分担青料费用和试验组（喂维生素添加剂）每只鸡分担的费用，进而计算出饲喂维生素添加剂的肉鸡全年可节约的费用。

（七）已具备的条件和研究进度安排 已具备的条件主要包括过去的研究工作基础或预试情况，现有的主要仪器设备，研究技术人员及协作条件，从其他渠道已得到的经费情况等。研究进度安排可根据试验的不同内容按日期、分阶段进行安排，定期写出总结报告。

（八）试验所需的条件 除已具备的条件外，本试验尚需的条件，如经费、饲料、仪器设备的数量和要求等。

（九）研究人员分工 一般分为主持人、主研人、参加人。在有条件的情况下，应以学历、职称较高并有丰富专业知识和实践经验的人员担任主持人或主研人，高、中、初级专业人员相结合，老、中、青相结合，使年限较长的研究项目能够后继有人，保持试验的连续性、稳定性和完整性。

（十）试验的时间，地点和工作人员 试验的时间，地点要安排合适，工作人员要固定，并参加一定培训，以保证试验正常进行。

（十一）成果鉴定及撰写学术论文 这是整个研究工作的最后阶段，凡属国家课题应召开鉴定会议，由同行专家作出评价。个人选择课题可以撰写学术论文发表自己的研究成果，根据试验结果作出理论分析，阐明事物在内在规律，并提出自己的见解和新的学术观点。一些重要的个人研究成果，也可以申请相关部门鉴定和国家专利。

二、试验方案的拟定

（一）试验方案的基本概念 试验方案（**experimental scheme**）是指根据试验目的与要求而拟定的进行比较的一组试验处理的总称。试验方案是整个试验工作的核心部分，因此，须周密考虑，慎重拟定。试验方案按供试因素的多少可区分为单因素试验方案、多因素试验方案。

1、单因素试验方案 单因素试验（**single-factor experiment**）是指整个试验中只比较一个试验因素的不同水平的试验。单因素试验方案由该试验因素的所有水平构成。这是最基本、最简单的试验方案。例如在猪饲料中添加 4 种剂量的土霉素，进行饲养试验。这是一个有 4 个水平的的单因素试验，添加土霉素的 4 种剂量，即该因素的 4 个水平就构成了试验方案。

2、多因素试验方案 多因素试验（**multiple-factor or factorial experiment**）是指在同一试验中同时研究两个或两个以上试验因素的试验。多因素试验方案由该试验的所有试验因素的水平组合（即处理）构成。多因素试验方案分为完全方案和不完全方案两类。

（1）完全方案 在列出因素水平组合（即处理）时，要求每一个因素的每个水平都要碰见一次，这时，水平组合（即处理）数等于各个因素水平数的乘积。例如以 3 种饲料配方对 3 个品种肉鸭进行试验。两个因素分别为饲料配方（A）、品种（B）。饲料配方（A）分为 A_1、A_2、A_3 水平，品种（B）分为 B_1、B_2、B_3 水平。共有 A_1B_1、A_1B_2、A_1B_3、A_2B_1、A_2B_2、A_2B_3、A_3B_1、A_3B_2、A_3B_3 共 3×3=9 个水平组合（处理）。这 9 个水平组合（处理）就构成了这两个因

素的试验方案。根据完全试验方案进行的试验称为全面试验。全面试验既能考察试验因素对试验指标的影响，也能考察因素间的交互作用，并能选出最优水平组合，从而能充分揭示事物的内部规律。多因素全面试验的效率高于多个单因素试验的效率。全面试验的主要不足是，当因素个数和水平数较多时，水平组合（处理）数太多，以至于在试验时，人力、物力、财力、场地等都难以承受，试验误差也不易控制。因而全面试验宜在因素个数和水平数都较少时应用。

（2）不完全方案　这也是一种多因素试验方案，但与上述多因素试验完全方案不同。它是将试验因素的某些水平组合在一起形成少数几个水平组合。这种试验方案的目的在于探讨试验因素中某些水平组合的综合作用，而不在于考察试验因素对试验指标的影响和交互作用。这种在全部水平组合中挑选部分水平组合获得的方案称为不完全方案。根据不完全方案进行的试验称为部分试验。动物试验的综合性试验（**comprehensive experiment**）、正交试验（**orthogonal experiment**）都属于部分试验。

综合性试验是针对起主导作用且相互关系已基本清楚的因素设置的试验，它的水平组合就是一系列经过实践初步证实的优良水平的配套。正交试验是在全部水平组合中选出有代表性的部分水平组合设置的试验，具体内容见第八节。

一个周密、完善的试验方案，不仅可以节省人力、物力，多快好省地完成试验任务，而且可以获得正确的试验结论。如果方案拟定不合理，如因素、水平选择不当，部分试验方案所包含的水平组合针对性或代表性差，试验将得不出应有的结果，甚至导致试验的失败。因此，试验方案的拟定在整个试验中占着极其重要的位置。

（二）拟定试验方案的要点　为了拟定一个正确的、切实可行的试验方案，应从以下几方面考虑：

1、根据试验的目的、任务和条件挑选试验因素　拟定方案时，在正确掌握生产中存在的问题后，对试验目的、任务进行仔细分析，抓住关键，突出重点。首先要挑选对试验指标影响较大的关键因素。若只考察一个因素，则可采用单因素试验。若是考察两个以上因素，则应采用多因素试验。如进行猪饲料添加某种微量元素的饲养试验，在拟定试验方案时，设置一个添加一定剂量微量元素的处理和不添加微量元素的对照，得到一个包含 2 个处理的单因素试验方案；或设置几个添加不同剂量微量元素的处理和一个不添加微量元素的对照，得到一个包含多个处理的单因素试验方案。进行微量元素不同添加剂量与不同品种猪的饲养试验，则安排一个二因素试验方案。应该注意，一个试验中研究的因素不宜过多，否则处理数太多，试验过于宠大，试验干扰因素难以控制。凡是能用简单方案的试验，就不用复杂方案。

2、根据各试验因素的性质分清水平间差异　各因素水平可根据不同课题、因素的特点及动物的反应能力来确定，以使处理的效应容易表现出来。

（1）水平的数目要适当　水平数目过多，不仅难以反映出各水平间的差异，而且加大了处理数；水平数太少又容易漏掉一些好的信息，至使结果分析不全面。

（2）水平间的差异要合理　有些因素在数量等级上只需少量的差异就反映出不同处理的效应。如饲料中微量元素的添加等。而有些则需较大的差异才能反应出不同处理效应来，如饲料用量等。

（3）试验方案中各因素水平的排列要灵活掌握　一般可采用等差法（即等间距法）、等比法和随机法 3 种。我们结合肉牛埋植玉米赤霉醇为例说明。

等差法是指各相邻两个水平数量之差相等，如赋形剂（不含玉米赤霉醇）各水平的排列

为：10mg、20mg、30mg，其中 20mg 为中心水平，向上向下都相隔 10mg。

等比法是指各相邻两个水平的数量比值相同，如赋形剂各水平的排列为 7.5mg、15mg、30mg、60mg,相邻两水平之比为 1:2。

随机法是指因素各水平随机排列，如赋形剂各水平排列为 15mg、10mg、40mg、30mg 各水平的数量无一定关系。

3、试验方案中必须设立作为比较标准的对照　动物试验的目的就是通过比较来鉴别处理效应大小、好坏等。为了达到这一目的，试验方案应当包括各试验处理，以及作为比较的对照。任何试验都不能缺少对照，否则就不能显示出试验的处理效果。根据研究的目的与内容，可选择不同的对照形式。例如，进行添加微量元素的饲养试验，添加微量元素为处理，不添加添加微量元素为对照，这样的对照为空白对照。进行几种微量元素添加量的比较试验，各个处理可互为对照，不必再设对照。在对某种动物作生理生化指标检验时，所得数据是否异常应与动物的正常值作比较,动物的正常值就是所谓的标准对照。在杂交试验中，要确定杂交优势的大小，必须以亲本作对照，这就是试验对照。另外，还有一种自身对照，即处理与对照在同一动物上进行，如病畜用药前与用药后生理指标的比较等。

4、试验处理（包括对照）之间应遵循唯一差异原则　这是指在进行处理间比较时，除了试验处理不同外，其它所有条件应当尽量一致或相同，使其具有可比性，才能使处理间的比较结果可靠。例如，进行不同品种猪的育肥比较试验，各参试猪除了品种不同外，其它如性别、年龄、体重等应一致，饲料和饲养管理等条件都应相同，才能准确评定品种的优劣。

5、有的试验要设置预试期　所谓预试就是正式试验开始之前根据试验设计进行的过渡试验，为正式试验做好准备工作。通过预饲，使供试的动物适应新的环境，对不合适的试验动物进行调整和淘汰，同时也使试验人员熟悉操作方法和程序。预试期的长短，可根据具体情况决定，一般以 10-20 天为宜。预试期间供试动物的数量应适当多于正式试验所需的数量。通过对预试所得到的数据资料的分析，还可检查试验设计的科学性、合理性和可行性，发现问题及时解决。

第三节　试验设计的基本原则

一、试验误差的来源

（一）试验误差　在畜牧、水产等科学研究中，试验处理常常受到各种非处理因素的影响，使试验处理的效应不能真实地反映出来，也就是说，试验所得到的观测值，不但有处理的真实效应，而且还包含其它因素的影响，这就出现了实测值与真值的差异，这种差异在数值上的表现称为试验误差。

由于产生误差的原因和性质不同，试验误差可分为系统误差（片面误差）、随机误差（抽样误差）两类。有关内容已在第一章第二节中详细阐述，这里不再重复。

（二）动物试验中误差的来源　系统误差影响试验的准确性，随机误差影响试验的精确性。为了提高试验的准确性与精确性，即提高试验的正确性，必须避免系统误差，降低随机误差。为了有效地避免系统误差，降低随机误差，必须了解试验误差的来源。动物试

验误差的主要来源有：

1、供试动物固有的差异　是指各处理的供试动物在遗传和生长发育上或多或少的差异性。如试验动物的遗传基础、性别、年龄、体重不同，生理状况、生产性能的不一致等，即使是全同胞间或同一个体不同时期间也会存在差异。

2、饲养管理不一致所引起的差异　指在试验过程中各个处理在饲养技术、管理方法及日粮配合等在质量上的不一致，以及在观测记载时由于工作人员的认真程度，掌握的标准不同或测量时间、仪器的不同等所引起的偏差。

3、环境条件的差异　主要指那些不易控制的环境的差异，如栏舍温度、湿度、光照、通风不同所引起的差异等。

4、由一些随机因素引起的偶然差异　如偶然疾病的侵袭、饲料的不稳定等引起的差异。

二、试验设计的基本原则

在动物试验中，误差主要是由于供试动物个体之间的差异和饲养管理不一致所造成。针对误差的主要来源，应采取切实有效的措施，如尽量选择初始条件一致的试验动物，尽量做到饲养管理一致，认真细致进行观测记载等，力求避免系统误差，降低随机误差。统计学上通过合理的试验设计既能获得试验处理效应与试验误差的无偏估计，也能控制和降低随机误差，提高试验的精确性。在试验设计时必须遵循以下基本原则。

（一）重复　重复是指试验中同一处理实施在两个或两个以上的试验单位上。在动物试验中，一头动物可以构成一个试验单位，有时一组动物也可构成一个试验单位。设置重复的主要作用在于估计试验误差和降低试验误差。如果同一处理只实施在一个试验单位上，那么只能得到一个观测值，则无从看出差异，因而无法估计试验误差的大小。只有当同一处理实施在两个或两个以上的试验单位上，获得两个或两个以上的观测值时，才能估计出试验误差。在第四章已经讲到，样本标准误与标准差的关系是 $S_{\bar{x}} = S/\sqrt{n}$，即平均数抽样误差的大小与重复次数的平方根成反比，故重复次数多可以降低试验误差。但在实际应用时，重复数太多，试验动物的初始条件不易控制一致，也不一定能降低误差。重复数的多少可根据试验的要求和条件而定。如果供试动物个体间差异较大，重复数应多些，差异较小，重复数可少些。

（二）随机化　随机化是指在对试验动物进行分组时必须使用随机的方法，使供试动物进入各试验组的机会相等，以避免试验动物分组时试验人员主观倾向的影响。这是在试验中排除非试验因素干扰的重要手段，目的是为了获得无偏的误差估计量。

（三）局部控制——试验条件的局部一致性　局部控制是指在试验时采取一定的技术措施或方法来控制或降低非试验因素对试验结果的影响。在试验中，当试验环境或试验单位差异较大时，仅根据重复和随机化两原则进行设计不能将试验环境或试验单位差异所引起的变异从试验误差中分离出来，因而试验误差大，试验的精确性与检验的灵敏度低。为解决这一问题，在试验环境或试验单位差异大的情况下，根据局部控制的原则，可将整个试验环境或试验单位分成若干个小环境或小组，在小环境或小组内使非处理因素尽量一致。每个比较一致的小环境或小组，称为单位组（或区组）。因为单位组之间的差异可在方差分析时从试验误差中分离出来，所以局部控制原则能较好地降低试验误差。

以上所述重复、随机化、局部控制三个基本原则称为费雪（**R.A.Fisher**）三原则，是试验设计中必须遵循的原则，再采用相应的统计分析方法，就能够最大程度地降低并无偏估计试验误差，无偏估计处理的效应，从而对于各处理间的比较作出可靠的结论。试验设计三原则的关系和作用见图 12-1 所示。

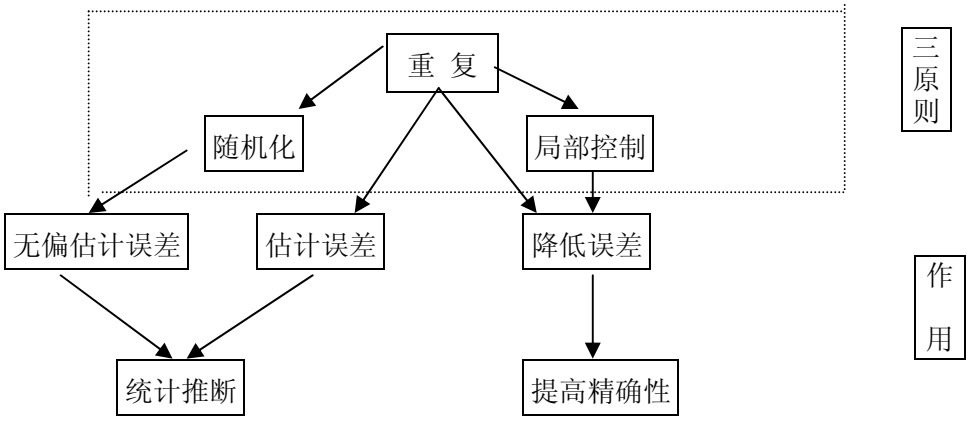


图 12-1　试验设计三原则的关系

第四节　完全随机设计

完全随机设计（**completely randomized design**）是根据试验处理数将全部供试动物随机地分成若干组，然后再按组实施不同处理的设计。这种设计保证每头供试验动物都有相同机会接受任何一种处理，而不受试验人员主观倾向的影响。在畜牧、水产等试验中，当试验条件特别是试验动物的初始条件比较一致时，可采用完全随机设计。这种设计应用了重复和随机化两个原则，因此能使试验结果受非处理因素的影响基本一致，真实反映出试验的处理效应。

完全随机设计的实质是将供试动物随机分组。随机分组的方法有抽签法和用随机数字表法，以用随机数字表法为好，因为随机数字表上所有的数字都是按随机抽样原理编制的，表中任何一个数字出现在任何一个位置都是完全随机的。除从随机数字表（见附表 13）可查得随机数字外，有些电脑及计算器均有此功能，用起来则更方便。下面举例说明用随机数字表将试验动物分组的方法。

一、完全随机的分组方法

（一）两个处理比较的分组

【例 12.1】 现有同品种、同性别、同年龄、体重相近的健康绵羊 18 只，试用完全随机的方法分成甲、乙两组。

绵羊编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
随机数字	16	07	44	99	83	11	46	32	24	20	14	85	88	45	10	93	72	88
组　别	乙	甲	乙	甲	甲	甲	乙	乙	乙	乙	乙	甲	乙	甲	乙	甲	乙	乙
调整组别							甲		甲									

首先将 18 只绵羊依次编为 1，2，……，18 号，然后从随机数字表中任意一个随机数字开始，向任一方向（左、右、上、下）连续抄下 18 个（两位）数字，分别代表 18 只绵羊。令随机数字中的单数为甲组，双数为乙组。如从随机数字表（Ⅰ）第 12 行第 7 列的 16 开始向右连续抄下 18 个随机数字填入表第二行。

随机分组结果：

甲组：2　4　5　6　12　14　16

乙组：1　3　7　8　9　10　11　13　15　17　18

甲组比乙组少 4 只，需要从乙组调整两只到甲组。仍用随机的方法进行调整。在前面 18 个随机数字后再接着抄下两个数字：71、23，分别除以 11（调整时乙组的绵羊只数）、10（调整 1 只绵羊去甲组后乙组剩余的绵羊只数），余数为 5、3，则把分配于乙组的第 5 只绵羊（9 号）和余下 10 只的第 3 只绵羊（7 号）分到甲组。调整后的甲、乙两组绵羊编号为：

甲组	2	4	5	6	7	9	12	14	16
乙组	1	3	8	10	11	13	15	17	18

（二）三个以上处理比较的分组

【例 12.2】 设有同品种、同性别、体重相近的健康仔猪 18 头，按体重大小依次编为 1、2、3、…、18 号，试用完全随机的方法，把它们等分成甲、乙、丙三组。

由随机数字表（Ⅱ）第 10 列第 2 个数 94 开始，向下依次抄下 18 个数，填入下表第 2 横行。

动物编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
随机数字	94	94	88	46	56	00	04	00	26	56	48	91	90	88	26	53	12	25
以3除后之余数	1	1	1	1	2	0	1	0	2	2	0	1	0	1	2	2	0	1
组　别	甲	甲	甲	甲	乙	丙	甲	丙	乙	乙	丙	甲	丙	甲	乙	乙	丙	甲
调整组别												丙		乙				

一律以 3（处理数）除各随机数字，若余数为 1，即将该动物归于甲组；余数为 2，归入乙组；商为 0 或余数为 0，归入丙组。结果归入甲组者 8 头，乙组 5 头，丙组 5 头。各组头数不等，应将甲组多余的 2 头调整 1 头给乙组、1 头给丙组。调整甲组的 2 头动物仍然采用随机的方法。从随机数字 25 后面接下去抄二个数 63、62，然后分别以 8（甲组原分配 8 头）、7 除之（注意：若甲组原分配有 9 头，须将多余的 3 头调整给另外两组，则抄下三个随机数，分别以 9、8、7 除之），得第一个余数为 7，第二个余数为 6，则把原分配在甲组的 8 头仔猪中第 7 头仔猪即 14 号仔猪改为乙组；把甲组中余下的 7 头仔猪中的第 6 头仔猪即 12 号仔猪改为丙组。这样各组的仔猪数就相等了。调整后各组的仔猪编号如下：

组　别	仔　猪　编　号					
甲　组	1	2	3	4	7	18
乙　组	5	9	10	14	15	16
丙　组	6	8	11	12	13	17

以上是用完全随机的方法，将试验动物分为两组或三组的情形，若将试验动物分为四组、五组或更多的组，方法相同。

二、试验结果的统计分析

对于完全随机试验的统计分析，由于试验处理数不同，统计分析方法也不同。

（一）处理数为 2 两个处理的完全随机设计也就是非配对设计，对其试验结果进行统计分析时，无论实际所得资料两处理重复数相同与否均采用非配对设计的 *t* 检验法分析。

（二）处理数大于 2 若获得的资料各处理重复数相等，则采用各处理重复数相等的单因素试验资料方差分析法分析；若在试验中，因受到条件的限制或供试动物出现疾病、死亡等使获得的资料各处理重复数不等，则采用各处理重复数不等的单因素试验资料方差分析法分析。

三、完全随机设计的优缺点

完全随机设计是一种最简单的设计方法，主要优缺点如下：

（一）完全随机设计的主要优点

1、设计容易 处理数与重复数都不受限制，适用于试验条件、环境、试验动物差异较小的试验。

2、统计分析简单 无论所获得的试验资料各处理重复数相同与否，都可采用 *t* 检验或方差分析法进行统计分析。

（二）完全随机设计的主要缺点

1、由于未应用试验设计三原则中的局部控制原则，非试验因素的影响被归入试验误差，试验误差较大，试验的精确性较低。

2、在试验条件、环境、试验动物差异较大时，不宜采用此种设计方法。

第五节 随机单位组设计

随机单位组设计（**randomized block design**）也称为随机区组（或窝组）设计。它是根据局部控制的原则，如将同窝、同性别、体重基本相同的动物划归一个单位组，每一单位组内的动物数等于处理数，并将各单位组的试验动物随机分配到各处理组，这种设计称为随机单位组设计。

随机单位组设计要求同一单位组内各头（只）试验动物尽可能一致，不同单位组间的试验动物允许存在差异，但每一单位组内试验动物的随机分组要独立进行，每种处理在一个单位组内只能出现一次。例如，为了比较 5 种不同中草药饲料添加剂对猪增重的效果，从 4 头母猪所产的仔猪中，每窝选出性别相同、体重相近的仔猪各 5 头，共 20 头，组成 4 个单位组，设计时每一单位组有仔猪 5 头，每头仔猪随机地喂给不同的饲料添加剂。这就是处理数为 5，单位组数为 4 的随机单位组设计。

一、随机单位组设计方法

（一）随机单位组设计的分组方法 在畜牧、水产等动物试验中，除把初始条件相同的动物如同窝仔畜划为同一单位组外，还可根据实际情况，把不同试验场、同一场内不同畜舍、不同池塘等划分为单位组。下面结合例子说明分组的方法。

【例 12.3】 前面提到的 5 种中草药饲料添加剂分别以 A_1、A_2、A_3、A_4、A_5表示，供试 4 窝仔猪分别按体重依次编号为：1-5 号为第Ⅰ组，6-10 号为第Ⅱ组，11-15 号为第Ⅲ组，16-20 为第Ⅳ组。试按随机单位组设计将试验仔猪分组。

表 12-2　5 种饲料添加剂试验随机单位组设计表

仔猪编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
随机数字	15	50	75	25	–	71	38	86	58	–	95	98	56	85	–	99	83	21	62	–
除　数	5	4	3	2	–	5	4	3	2	–	5	4	3	2	–	5	4	3	2	–
余　数	5	2	3	1	–	1	2	2	2	–	5	2	2	1	–	4	3	3	2	–
添加剂	A_5	A_2	A_4	A_1	A_3	A_1	A_3	A_4	A_5	A_2	A_5	A_2	A_3	A_1	A_4	A_4	A_3	A_5	A_2	A_1

先从随机数字表（Ⅱ）第 15 行、第 11 列 15 开始，向下依次抄下 16 个随机数字（舍弃 00），每抄 4 个数字留一空位，见表 12-2 第 2 行。再将同一单位组内前 4 个随机数字依次除以 5、4、3、2（最大数 5 为处理数），根据余数（余数为 0 者，以除数代之）确定每一单位组内各供试仔猪喂给的添加剂种类。如第一单位组中，第一个余数是 5，则将第 1 号仔猪喂给 5 种添加剂列于第 5 位的 A_5添加剂；第二个余数是 2，则将第 2 号仔猪喂给剩下的 4 种添加剂 A_1、A_2、A_3、A_4列于第二位的 A_2添加剂；第三个余数是 3，则将第 3 号仔猪喂给剩下的 3 种添加剂 A_1、A_3、A_4列于第三位的 A_4添加剂；第四个余数是 1，则将第 4 号仔猪喂给剩下的 2 种添加剂 A_1、A_3列于第 1 位的 A_1添加剂；第 5 号仔猪只能喂给剩下的 A_3添加剂。用同样方法一一确定其它单位组内各仔猪喂给的添加剂，结果见表 12-3。

表 12-3　5 种饲料添加剂试验随机单位组设计试验动物分组表

添加剂	单　位　组			
	Ⅰ	Ⅱ	Ⅲ	Ⅳ
A_1	4	6	14	20
A_2	2	10	12	19
A_3	5	7	13	17
A_4	3	8	15	16
A_5	1	9	11	18

（二）配对设计分组方法　配对设计是处理数为 2 的随机单位组设计。在进行配对设计时，配成对子的两个试验单位必须符合配对要求：配成对子的两个试验单位的初始条件尽量一致，不同对子间试验单位的初始条件允许有差异，每一个对子就是试验处理的一个重复，然后将配成对子的两个试验单位随机地分配到两个处理组中。

例如，现有同一品种的供试家畜 18 头，分别将性别、年龄相同，体重相似的两头家畜配成对子，共 9 对，编号为 1-9 号。试用随机方法将每个对子中的两头家畜分到甲、乙两个处理组中。

由随机数字表（Ⅰ）（附表 13）的第 16 行、第 8 列 20 开始，向右依次抄下 9 个随机数字，将单数组中配对的第一头家畜归入甲组，第二头家畜归入乙组；双数组中配对的第一头家畜归入乙组，第二头家畜归入甲组，则 9 对家畜分组如下：

配对编号	1	2	3	4	5	6	7	8	9
随机数字	20	38	26	13	89	51	03	74	17
配对中第一头家畜组别	乙	乙	乙	甲	甲	甲	甲	乙	甲
配对中第二头家畜组别	甲	甲	甲	乙	乙	乙	乙	甲	乙

二、试验结果的统计分析

（一）随机单位组试验结果的统计分析 随机单位组试验结果的统计分析采用方差分析法。分析时将单位组也看成一个因素，连同试验因素一起，按两因素单独观测值的方差分析法进行。这里需要说明的是，假定单位组因素与试验因素不存在交互作用。

若记试验处理因素为 A，处理因素水平数为 a；单位组因素为 B，单位组数为 b，对试验结果进行方差分析的数学模型为：

$$x_{ij}=\mu+\alpha_i+\beta_j+\varepsilon_{ij} \quad (i=1,2,\ldots,a;j=1,2,\ldots b) \qquad (12\text{-}1)$$

式中 μ 为总体均数，α_i 为第 i 处理的效应，β_j 为第 j 单位组效应。处理效应 α_i 通常是固定的，且有 $\sum_{i=1}^{a}\alpha_i=0$ ；单位组效应 β_j 通常是随机的。ε_{ij} 为随机误差，相互独立，且都服从 $N(0,\sigma^2)$ 。

平方和与自由度的划分式为：

$$\begin{aligned} SS_T &= SS_A+SS_B+SS_e \\ df_T &= df_A+df_B+df_e \end{aligned} \qquad (12\text{-}2)$$

对于**【例 12.3】**，通过按表 12-3 试验动物分组结果进行试验后，各号仔猪增重结果列于表 12-4。

表 12-4　5 种不同饲料添加剂对仔猪的增重效果　　（单位：g）

处　理(A)	单　位　组(B)				处理合计	处理平均
	B_{I}	B_{II}	B_{III}	B_{IV}	$x_{i.}$	$\bar{x}_{i.}$
A_1	205	168	222	230	825	206.25
A_2	230	198	242	255	925	231.25
A_3	252	248	305	260	1065	266.25
A_4	200	158	183	196	737	184.25
A_5	265	275	315	282	1137	284.25
单位组合计 $x_{.j}$	1152	1047	1267	1223	4689($x_{..}$)	

1、计算各项平方和与自由度

矫正数　$C=x_{..}^2/ab=4689^2/5\times4=1099336.05$

总平方和　$SS_T=\sum x_{ij}^2-C=(205^2+168^2+\cdots+282^2)-1099336.06=35890.95$

处理间平方和　$SS_A=\sum x_{i.}^2/b-C=(825^2+925^2+\cdots+1137^2)/4-1099336.05=27267.2$

单位组间平方和　$SS_B=\sum x_{.j}^2/a-C=(1152^2+1047^2+\cdots+1223^2)/5-1099336.05=5530.15$

误差平方和　$SS_e=SS_T-SS_A-SS_B=35890.95-27267.2-5530.15=3093.6$

总自由度　$df_T=ab-1=5\times4-1=19$

处理间自由度　$df_A=a-1=5-1=4$

单位组间自由度　$df_B=b-1=4-1=3$

误差自由度　$df_e=df_T-df_A-df_B=(a-1)(b-1)=(5-1)(4-1)=12$

2、列出方差分析表，进行 F 检验

表 12-5　方差分析表

变异原因	SS	df	MS	F	$F_{0.01}$
处 理 间（A）	27267.20	4	6816.80	26.44^{**}	5.41
单位组间（B）	5530.15	3	1843.38	7.15^{**}	5.95
误　　差	3093.6	12	257.8		
总 变 异	35890.95	19			

因为 $F_A > F_{0.01(4,12)}$，$F_B > F_{0.01(3,12)}$，表明饲料添加剂对仔猪增重影响极显著，因而还需要对各不同饲料添加剂平均数间差异的显著性进行检验。单位组间的变异，虽然 F 值已达到 0.01 显著水平，由于我们采取的是随机单位组设计，已将它从误差中分离出来，达到了局部控制的目的。单位组间的变异即使显著，一般也不作单位组间的多重比较。

3、饲料添加剂间的多重比较

表 12-6　饲料添加剂平均数间多重比较表（q 法）

添加剂	平均数 $\bar{x}_i.$	$\bar{x}_i.-184.25$	$\bar{x}_i.-206.25$	$\bar{x}_i.-231.25$	$\bar{x}_i.-266.25$
A_5	284.25	100^{**}	78^{**}	53^{**}	18
A_3	266.25	82^{**}	60^{**}	35^{**}	
A_2	231.25	47^{**}	25^{*}		
A_1	206.25	22			
A_4	184.25				

均数标准误为：

$$S_{\bar{x}} = \sqrt{MS_e / n} = \sqrt{257.8/4} = 8.028$$

由 df_e=12、秩次距 k=2，3，4，5，查附表 5 得临界 q 值：$q_{0.05}$、$q_{0.01}$，并与 $S_{\bar{x}}$ 相乘求得 LSR 值，列于表 12-7。

表 12-7　q 值和 LSR 值表

df_e	k	$q_{0.05}$	$q_{0.01}$	$LSR_{0.05}$	$LSR_{0.01}$
12	2	3.08	4.32	24.73	34.68
	3	3.77	5.04	30.27	40.46
	4	4.20	5.50	33.72	44.15
	5	4.51	5.84	36.21	46.88

由表 12-6 看出，除 A_5 与 A_3，A_1 与 A_4 之间差异不显著，A_2 与 A_1 间差异显著外，其余平均数间差异极显著，说明采用 A_5、A_3 添加剂仔猪平均增重极显著高于 A_2、A_1、A_4 添加剂；A_2 显著高于 A_1、极显著高于 A_4；A_4 添加剂对仔猪增重效果最差。

（二）配对设计试验结果的统计分析　试验结果为计量资料时，采用第五章所介绍的配对设计 t 检验法进行统计分析。若试验结果为次数资料，采用配对次数资料的 χ^2 检验法进行分析。

三、随机单位组设计的优缺点

（一）随机单位组设计的主要优点

1、设计与分析方法简单易行。

2、由于随机单位组设计体现了试验设计三原则，在对试验结果进行分析时，能将单位组间的变异从试验误差中分离出来，有效地降低了试验误差，因而试验的精确性较高。

3、把条件一致的供试动物分在同一单位组，再将同一单位组的供试动物随机分配到不同处理组内，加大了处理组之间的可比性。

（二）随机单位组设计的主要缺点 当处理数目过多时，各单位组内的供试动物数数目也过多，要使各单位组内供试动物的初始条件一致将有一定难度，因而在随机单位组设计中，处理数以不超过 20 为宜。

配对设计是处理数为 2 的随机单位组设计，其优点是结果分析简单，试验误差通常比非配对设计小，但由于试验动物配对要求严格，不允许将不满足配对要求的试验动物随意配对。

第六节 拉丁方设计

“拉丁方”的名字最初是由 R、A、Fisher 给出的。拉丁方设计（**latin square design**）是从横行和直列两个方向进行双重局部控制，使得横行和直列两向皆成单位组，是比随机单位组设计多一个单位组的设计。在拉丁方设计中，每一行或每一列都成为一个完全单位组，而每一处理在每一行或每一列都只出现一次，也就是说，在拉丁方设计中，试验处理数=横行单位组数=直列单位组数=试验处理的重复数。在对拉丁方设计试验结果进行统计分析时，由于能将横行、直列二个单位组间的变异从试验误差中分离出来，因而拉丁方设计的试验误差比随机单位组设计小，试验精确性比随机单位组设计高。

一、拉丁方简介

（一）拉丁方 以 *n* 个拉丁字母 *A*，*B*，*C*……，为元素，作一个 *n* 阶方阵，若这 *n* 个拉丁方字母在这 *n* 阶方阵的每一行、每一列都出现、且只出现一次，则称该 *n* 阶方阵为 *n*×*n* 阶拉丁方。

例如：

A *B*	*B* *A*
B *A*	*A* *B*

为 2×2 阶拉丁方，2×2 阶拉丁方只有这两个。

A	*B*	*C*
B	*C*	*A*
C	*A*	*B*

为 3×3 阶拉丁方。

第一行与第一列的拉丁字母按自然顺序排列的拉丁方，叫标准型拉丁方。3×3 阶标准型拉丁方只有上面介绍的 1 种，4×4 阶标准型拉丁方有 4 种，5×5 阶标准型拉丁方有 56 种。若变换标准型的行或列，可得到更多种的拉丁方。在进行拉丁方设计时，可从上述多种拉丁方中随机选择一种；或选择一种标准型，随机改变其行列顺序后再使用。

（二）常用拉丁方 在动物试验中，最常用的有 3×3，4×4，5×5，6×6 阶拉丁方。下面列出部分标准型拉丁方，供进行拉丁方设计时选用。其余拉丁方可查阅数理统计表及有关参考书。

3×3

A	*B*	*C*
B	*C*	*A*
C	*A*	*B*

4 × 4

（1）

A	*B*	*C*	*D*
B	*A*	*D*	*C*
C	*D*	*B*	*A*
D	*C*	*A*	*B*

（2）

A	*B*	*C*	*D*
B	*C*	*D*	*A*
C	*D*	*A*	*B*
D	*A*	*B*	*C*

（3）

A	*B*	*C*	*D*
B	*D*	*A*	*C*
C	*A*	*D*	*B*
D	*C*	*B*	*A*

（4）

A	*B*	*C*	*D*
B	*A*	*D*	*C*
C	*D*	*A*	*B*
D	*C*	*B*	*A*

5 × 5

（1）

A	*B*	*C*	*D*	*E*
B	*A*	*E*	*C*	*D*
C	*D*	*A*	*E*	*B*
D	*E*	*B*	*A*	*C*
E	*C*	*D*	*B*	*A*

（2）

A	*B*	*C*	*D*	*E*
B	*A*	*D*	*E*	*C*
C	*E*	*B*	*A*	*D*
D	*C*	*E*	*B*	*A*
E	*D*	*A*	*C*	*B*

（3）

A	*B*	*C*	*D*	*E*
B	*A*	*E*	*C*	*D*
C	*E*	*D*	*A*	*B*
D	*C*	*B*	*E*	*A*
E	*D*	*A*	*B*	*C*

（4）

A	*B*	*C*	*D*	*E*
B	*A*	*D*	*E*	*C*
C	*D*	*E*	*A*	*B*
D	*E*	*B*	*C*	*A*
E	*C*	*A*	*B*	*D*

6 × 6

A	*B*	*C*	*D*	*E*	*F*
B	*F*	*D*	*C*	*A*	*E*
C	*D*	*E*	*F*	*B*	*A*
D	*A*	*F*	*E*	*C*	*B*
E	*C*	*A*	*B*	*F*	*D*
F	*E*	*B*	*A*	*D*	*C*

二、拉丁方设计方法

在畜牧、水产等动物试验中，如果要控制来自两个方面的系统误差，且试验动物的数量又较少，则常采用拉丁方设计。下面结合具体例子说明拉丁方设计方法。

【例 12.4】 为了研究 5 种不同温度对蛋鸡产蛋量的影响，将 5 栋鸡舍的温度设为 *A*、*B*、*C*、*D*、*E*，把各栋鸡舍的鸡群的产蛋期分为 5 期，由于各鸡群和产蛋期的不同对产蛋量有较大的影响，因此采用拉丁方设计，把鸡群和产蛋期作为单位组设置，以便控制这两个方面的系统误差。拉丁方设计步骤如下：

（一）选择拉丁方 选择拉丁方时应根据试验的处理数和横行、直列单位组数先确定采用几阶拉丁方，再选择标准型拉丁方或非标准型拉丁方。此例因试验处理因素为温度，处理数为 5；将鸡群作为直列单位组因素，直列单位组数为 5；将产蛋期作为横行单位组因素，横行单位组数亦为 5，即试验处理数、直列单位组数、横行单位组数均为 5，则应选取 5×5 阶拉丁方。本例选取前面列出的第 2 个 5 × 5 标准型拉丁方，即：

A	*B*	*C*	*D*	*E*
B	*A*	*D*	*E*	*C*
C	*E*	*B*	*A*	*D*

D C E B A
E D A C B

（二）随机排列 在选定拉丁方之后，如是非标准型时，则可直接按拉丁方中的字母安排试验方案。若是标准型拉丁方，还应按下列要求对横行、直列和试验处理的顺序进行随机排列。

3×3 标准型拉丁方：直列随机排列，再将第二和第三横行随机排列。

4×4 标准型拉丁方：随机选择 4 个标准型拉丁方中的一个，然后再将横行、直列及处理都随机排列。

下面对选定的 5×5 标准型拉丁方进行随机排列。先从随机数字表（Ⅰ）第 22 行、第 8 列 97 开始，向右连续抄录 3 个 5 位数，抄录时舍去“0”、“6 以上的数”和重复出现的数，抄录的 3 个五位数字为：13542，41523，34521。然后将上面选定的 5×5 拉丁方的直列、横行及处理按这 3 个五位数的顺序重新随机排列。

1、直列随机 将拉丁方的各直列顺序按 13542 顺序重排。

2、横行随机 再将直列重排后的拉丁方的各横行按 41523 顺序重排。

选择拉丁方					直列随机						横行随机					
1	2	3	4	5		1	3	5	4	2						
A	*B*	*C*	*D*	*E*	1	*A*	*C*	*E*	*D*	*B*	4	*D*	*E*	*A*	*B*	*C*
B	*A*	*D*	*E*	*C*	2	*B*	*D*	*C*	*E*	*A*	1	*A*	*C*	*E*	*D*	*B*
C	*E*	*B*	*A*	*D*	3	*C*	*B*	*D*	*A*	*E*	5	*E*	*A*	*B*	*C*	*D*
D	*C*	*E*	*B*	*A*	4	*D*	*E*	*A*	*B*	*C*	2	*B*	*D*	*C*	*E*	*A*
E	*D*	*A*	*C*	*B*	5	*E*	*A*	*B*	*C*	*D*	3	*C*	*B*	*D*	*A*	*E*

3、把 5 种不同温度按第三个 5 位数 34521 顺序排列 即：*A*=3，*B*=4，*C*=5，*D*=2，*E*=1，也就是说，在拉丁方中的 *A* 表示第 3 种温度，*B* 表示第 4 种温度等，依次类推。从而得出 5×5 拉丁方设计，如表 12-8 所示。

表 12-8 5 种不同温度对鸡产蛋量影响的拉丁方设计

产蛋期	鸡群				
	一	二	三	四	五
Ⅰ	*D*（2）	*E*（1）	*A*（3）	*B*（4）	*C*（5）
Ⅱ	*A*（3）	*C*（5）	*E*（1）	*D*（2）	*B*（4）
Ⅲ	*E*（1）	*A*（3）	*B*（4）	*C*（5）	*D*（2）
Ⅳ	*B*（4）	*D*（2）	*C*（5）	*E*（1）	*A*（3）
Ⅴ	*C*（5）	*B*（4）	*D*（2）	*A*（3）	*E*（1）

注：括号内的数字表示温度的编号

由表 12-8 可以看出，第一鸡群在第Ⅰ个产蛋期用第 2 种温度，第二鸡群在第Ⅰ个产蛋期用第 1 种温度，等等。试验应严格按设计实施。

三、试验结果的统计分析

拉丁方设计试验结果的分析，是将两个单位组因素与试验因素一起，按三因素试验单独观测值的方差分析法进行，但应假定 3 个因素之间不存在交互作用。将横行单位组因素记为 *A*，直列单位组因素记为 *B*，处理因素记为 *C*，横行单位组数、直列单位组数与处理数记

为 r，对拉丁方试验结果进行方差分析的数学模型为：

$$x_{ij(k)}=\mu+\alpha_i+\beta_j+\gamma_{(k)}+\varepsilon_{ij(k)} \quad (i=j=k=1，2，\cdots，r) \qquad (12\text{-}3)$$

式中：μ 为总平均数；α_i 为第 i 横行单位组效应；β_j 为第 j 直列单位组效应，$\gamma_{(k)}$ 为第 k 处理效应。单位组效应α_i、β_j通常是随机的，处理效应$\gamma_{(k)}$通常是固定的，且有 $\sum_{k=1}^{\gamma}\gamma_{(k)}=0$；$\varepsilon_{ij(k)}$为随机误差，相互独立，且都服从 N（0，σ^2）。

注意：k 不是独立的下标，因为 i、j 一经确定，k 亦随之确定。

平方和与自由度划分式为：

$$SS_T=SS_A+SS_B+SS_C+SS_e$$
$$df_T=df_A+df_B+df_c+df_e \qquad (12\text{-}4)$$

【例 12.4】的试验结果如表 12-9 所示。

表 12-9　5 种不同温度对母鸡产蛋量影响试验结果　　（单位：个）

产蛋期	鸡群					横行和 $x_{i.}$
	一	二	三	四	五	
Ⅰ	*D*（23）	*E*（21）	*A*（24）	*B*（21）	*C*（19）	108
Ⅱ	*A*（22）	*C*（20）	*E*（20）	*D*（21）	*B*（22）	105
Ⅲ	*E*（20）	*A*（25）	*B*（26）	*C*（22）	*D*（23）	116
Ⅳ	*B*（25）	*D*（22）	*C*（25）	*E*（21）	*A*（23）	116
Ⅴ	*C*（19）	*B*（20）	*D*（24）	*A*（22）	*E*（19）	104
直列和 $x._j$	109	108	119	107	106	$x..$=549

注：括号内数字为产蛋量

表 12-10　各种温度（处理）的合计

温度	*A*	*B*	*C*	*D*	*E*
$x_{(k)}$	116	114	105	113	101
$\bar{x}_{(k)}$	23.2	22.8	21.0	22.6	20.2

现对表 12-9 资料进行方差分析。

1、计算各项平方和与自由度

矫正数　$C=x^2../r^2=549^2/5^2=12056.04$

总平方和　$SS_T=\Sigma x^2_{ij(k)}-C=23^2+21^2+\cdots\cdots+19^2-12056.04=12157-12056.04=100.96$

横行平方和　$SS_A=\Sigma x^2_{i.}/r-C=(108^2+105^2+\cdots\cdots+104^2)/5-12056.04=27.36$

直列平方和　$SS_B=\Sigma x^2_{.j}/r-C=(109^2+108^2+\cdots\cdots+106^2)/5-12056.04=22.16$

处理平方和　$SS_C=\Sigma x^2_{(K)}/r-C=(116^2+114^2+\cdots\cdots+101^2)/5-12056.04=33.36$

误差平方和　$SS_e=SS_T-SS_A-SS_B-SS_c=100.96-33.36-27.36-22.16=18.08$

总自由度　$df_T=r^2-1=5^2-1=24$

横行自由度　$df_A=r-1=5-1=4$

直列自由度　$df_B=r-1=5-1=4$

处理自由度　$df_C=r-1=5-1=4$

误差自由度　$df_e=df_T-df_A-df_B-df_C=(r-1)(r-2)=(5-1)(5-2)=12$

2、列出方差分析表，进行 F 检验

表 12-11　表 12-9 资料的方差分析表

变异来源	SS	df	MS	F	$F_{0.05}$	$F_{0.01}$
横行间	27.36	4	6.84	4.56^{*}	3.26	5.41
直列间	22.16	4	5.54	3.69^{*}	3.26	5.41
温度间	33.36	4	8.34	5.56^{**}	3.26	5.41
误　差	18.08	12	1.50			
总变异	100.96	24				

经 F 检验，产蛋期间和鸡群间差异显著,温度间差异极显著。因在拉丁方设计中，横行、直列单位组因素是为了控制和降低试验误差而设置的非试验因素，所以即使显著一般也不对单位组间进行多重比较。下面对不同温度平均产蛋量间作进行多重比较。

3、多重比较

列出多重比较表，见表 12-12。

表 12-12　不同温度平均产蛋量多重比较表（q 法）

温度	平均数 $\bar{x}_{(k)}$	$\bar{x}_{(k)}-20.2$	$\bar{x}_{(k)}-21$	$\bar{x}_{(k)}-22.6$	$\bar{x}_{(k)}-22.8$
A	23.2	3.0*	2.2	0.6	0.4
B	22.8	2.6*	1.8	0.2	
D	22.6	2.4*	1.6		
C	21.0	0.8			
E	20.2				

温度平均数标准误为:

$$S_{\bar{x}}=\sqrt{MS_e/n}=\sqrt{1.5/5}=0.55$$

由 df_e=12 和 k=2，3，4，5 从 q 值表查得临界 q 值：$q_{0.05}$ 和 $q_{0.01}$,并与 $S_{\bar{x}}$ 相乘得 LSR_{α} 值，列于表 12-13。

表 12-13　q 值和 LSR 值表

df_e	k	$q_{0.05}$	$q_{0.01}$	$LSR_{0.05}$	$LSR_{0.01}$
12	2	3.08	4.32	1.69	2.38
	3	3.77	5.04	2.07	2.77
	4	4.20	5.50	2.31	3.03
	5	4.51	5.84	2.48	3.21

多重比较结果表明：温度 A、B、D 平均产蛋量显著地高于 E，即第 3、4、2 种温度的平均产蛋量显著高于第 1 种温度的平均产蛋量，其余之间差异不显著。第 1 种和第 5 种温度平均产蛋量最低。

四、拉丁方设计的优缺点

（一）拉丁方设计的主要优点

1、精确性高　拉丁方设计在不增加试验单位的情况下，比随机单位组设计多设置了一个单位组因素，能将横行和直列两个单位组间的变异从试验误差中分离出来，因而试验误差比随机单位组设计小，试验的精确性比随机单位组设计高。

2、试验结果的分析简便

（二）拉丁方设计的主要缺点　因为在拉丁设计中，横行单位组数、直列单位组数、试验处理数与试验处理的重复数必须相等，所以处理数受到一定限制。若处理数少，则重复数也少，估计试验误差的自由度就小，影响检验的灵敏度；若处理数多，则重复数也多，横行、直列单位组数也多，导致试验工作量大，且同一单位组内试验动物的初始条件亦难控制一致。因此，拉丁方设计一般用于 5-8 个处理的试验。在采用 4 个以下处理的拉丁方设计时，为了使估计误差的自由度不少于 12，可采用“复拉丁方设计”，即同一个拉丁方试验重复进行数次，并将试验数据合并分析，以增加误差项的自由度。

应当注意，在进行拉丁方试验时，某些单位组因素，如奶牛的泌乳阶段，试验因素的各处理要逐个地在不同阶段实施，如果前一阶段有残效，在后一阶段的试验中，就会产生系统误差而影响试验的准确性。此时应根据实际情况，安排适当的试验间歇期以消除残效。另外，还要注意，横行、直列单位组因素与试验因素间不存在交互作用，否则不能采用拉丁方设计。

*第七节　交叉设计

交叉设计亦称反转试验设计，是指在同一试验中将试验单位分期进行、交叉反复二次以上的试验设计方法。

在动物试验中，为了提高试验的精确性，要求选用在遗传及生理上相同或相似的试验动物，但这在实践中往往不易满足。如进行奶牛的泌乳试验时，要选择若干头品种、性别、年龄、胎次等条件都相同的奶牛是很困难的。为了较好地消除试验动物个体之间以及试验时期间的差异对试验结果的影响，可采用交叉设计法。常用的有 2×2 和 2×3 交叉设计，见表 12-14 和 12-15。

表 12-14　2×2 交叉设计

群别	时　期	
	I	II
1	处理	对照
2	对照	处理

表 12-15　2×3 交叉设计

群别	时　　期		
	I	II	III
1	处理	对照	处理
2	对照	处理	对照

一、2×2 交叉设计与分析

2×2 交叉设计就是两组试验动物分两期一次交叉的试验设计。下面举例说明试验结果的分析方法。

【例 12.5】为了研究饲料新配方对奶牛产奶量的影响，设置对照饲料 A_1 和新饲料配方 A_2 两个处理，选择条件相近的奶牛 10 头，随机分为 B_1、B_2 两组，每组 5 头，预试期 1 周。试验分为 C_1、C_2 两期，每期两周，按 2×2 交叉设计进行试验。试验结果列于表 12-16。试

检验新饲料配方对提高产奶量有无效果。

对于 2×2 交叉试验资料，采用单因素二水平差值 d 的方差分析法**(Lucas)**或 t 检验法(明道绪)进行分析。两种分析方法的无效假设、备择假设均为：$Ho:\mu_{d_1}=\mu_{d_2}, H_A:\mu_{d_1}\neq\mu_{d_2}$。

（一）方差分析法 此例处理数 k=2，重复数 r=5。先计算出两个时期产奶量的差 $d=C_1-C_2$，以及 $T_1=\sum d_1$，$T_2=\sum d_2$，见表 12-16。

1、计算各项平方和与自由度

矫正数 $$C=(T_1+T_2)^2/kr=(-11.4+13.2)^2/(2\times5)=0.3240$$

总平方和 $$SS_T=\sum\sum d_{ij}^2-C=(-1.7)^2+(-2.2)^2+\cdots+1^2-0.3240$$
$$=75.4400-0.3240=75.1160$$

处理平方和 $$SS_A=\sum T_i^2/r-C=[(-11.4)^2+13.2^2]/5-0.3240=60.5160$$

误差平方和 $$SS_e=SS_T-SS_A=75.1160-60.5160=14.6000$$

总自由度 $$df_T=Kr-1=2\times5-1=9$$

处理自由度 $$df_A=K-1=2-1=1$$

误差自由度 $$df_e=df_T-df_A=k(r-1)=2(5-1)=8$$

表 12-16 【例 12.5】试验结果 (单位：千克/头·日)

时　　期		C_1	C_2	$d=C_1-C_2$	
处　　理		A_1	A_2	d_1	d_2
B_1	B_{11}	13.8	15.5	-1.7	
	B_{12}	16.2	18.4	-2.2	
	B_{13}	13.5	16.0	-2.5	
组	B_{14}	12.8	15.8	-3.0	
	B_{15}	12.5	14.5	-2.0	
处　　理		A_2	A_1		
B_2	B_{21}	14.3	13.5		0.8
	B_{22}	20.2	15.4		4.8
	B_{23}	18.6	14.3		4.3
组	B_{24}	17.5	15.2		2.3
	B_{25}	14.0	13.0		1.0
总　　和				T_1=-11.4	T_2=13.2

2、列出方差分析表，进行 F 检验

表 12-17 【例 12.5】试验资料方差分析表

变异来源	SS	df	MS	F	$F_{0.01(1,8)}$
处　理	60.5160	1	60.52	33.16**	11.26
误　差	14.6000	8	1.83		
总变异	75.116	9			

因为处理 F 值 33.16>$F_{0.01(1,8)}$，P<0.01，否定 $H_o:\mu_{d_1}=\mu_{d_2}$，接受 $H_A:\mu_{d_1}\neq\mu_{d_2}$，表明新配方饲料与对照饲料平均产奶量差异极显著，这里表现为新配方饲料的平均产奶量极显著高于对照饲料的平均产奶量。

（二）t 检验法 检验公式为：

$$t=\frac{\bar{d}_1-\bar{d}_2}{S_{\bar{d}_1-\bar{d}_2}},\qquad df=(r-1)+(s-1) \tag{12-5}$$

其中：$S_{\bar{d}_1-\bar{d}_2}=\sqrt{\dfrac{[\sum d_1^2-(\sum d_1)^2/r]+[\sum d_2^2-(\sum d_2)^2/s]}{(r-1)+(s-1)}\left(\dfrac{1}{r}+\dfrac{1}{s}\right)}$

——差数平均数差异标准误

r, s 分别为两组试验个体数。

此例，r=s=5，　$\bar{d}_1$ = -2.28，　$\bar{d}_2$ = 2.64

$$S_{\bar{d}_1-\bar{d}_2}=\sqrt{\frac{[(-1.7)^2+(-2.2)^2+\cdots+(-2.0)^2-\frac{(-11.4)^2}{5}]+[0.8^2+4.8^2+\cdots+1.0^2-\frac{13.2^2}{5}]}{(5-1)+(5-1)}\left(\frac{1}{5}+\frac{1}{5}\right)}$$

$$=0.8544$$

$$t=\frac{\bar{d}_1-\bar{d}_2}{S_{\bar{d}_1-\bar{d}_2}}=\frac{-2.28-2.64}{0.8544}=-5.7584$$

由 df=(r-1)+(s-1)=(5-1)+(5-1)=8 查临界 t 值得：$t_{0.01(8)}$=3.355，因为|t|=5.7854> $t_{0.01(8)}$，P<0.01，否定 $Ho:\mu_{d_1}=\mu_{d_2}$，接受 $H_A:\mu_{d_1}\neq\mu_{d_2}$，表明新配方饲料与对照饲料平均产奶量差异极显著。检验结果与方差分析法一致。

二、2×3 交叉设计与分析

2×3 交叉设计就是将试验动物分三期两次交叉的试验设计。对于 2×3 交叉试验资料，亦采用方差分析法或 t 检验法进行分析。

【例 12.6】 为了研究饲喂尿素对奶牛产奶量的影响，设置尿素配合饲料 A_1 和对照饲料 A_2 两个处理，选择条件相近的奶牛 6 头，随机分为 B_1、B_2 两组，每组 3 头，试验分 C_1、C_2、C_3 三期(每期 20 天)，B_1 组(B_{11}，B_{12}，B_{13})按 A_1—A_2—A_1 顺序给予饲料，B_2 组(B_{21}，B_{22}，B_{23})按 A_2—A_1—A_2 顺序给予饲料，预饲期 1 周。试验结果列于表 12-8，试检验尿素对提高奶牛的产奶量有无效果。

按公式 $d=C_1-2C_2+C_3$ 分别计算出 B_1 组的差 d_1 和 B_2 组的差 d_2 及 T_1、T_2。两种方法的无效假设与备择假设均为 $H_o:\mu_{d_1}=\mu_{d_2}$，$H_A:\mu_{d_1}\neq\mu_{d_2}$。

（一）方差分析法 此例，处理数 k=2，重复数 r=3。

1、计算各项平方和与自由度

矫正数　　$C=(T_1+T_2)^2/kr=(3.06-0.11)^2/6=1.4504$

总平方和　　$SS_T=\sum d^2-C=(-0.09)^2+0.70^2+\cdots+0.26^2-1.4504=5.8475$

处理平方和　$SS_A=(T_1^2+T_2^2)/r-C=[3.06^2+(-0.11)^2]/3-1.4504=1.6748$

误差平方和　$SS_e=SS_T-SS_A=5.8475-1.6748=4.1727$

总自由度　$df_T=kr-1=2\times3-1=5$

处理自由度　$df_A=k-1=2-1=1$

误差自由度　$df_e=k(r-1)=df_T-df_A=5-1=4$

表 12-18　【例 12.6】 试验结果　(单位：千克/头·日)

时　期		C_1	C_2	C_3	$d=C_1-2C_2-C_3$	
处　理		A_1	A_2	A_1	d_1	d_2
B_1	B_{11}	11.32	11.36	11.31	-0.09	
	B_{12}	13.67	13.40	13.83	0.70	
组	B_{13}	18.74	16.34	16.39	2.45	
处　理		A_2	A_1	A_2		
B_2	B_{21}	11.65	11.19	11.12		0.39
	B_{22}	13.57	13.87	13.41		-0.76
组	B_{23}	11.54	10.97	10.66		0.26
总　和					T_1=3.06	T_2=-0.11

2、列出方差分析表，进行 F 检验

表 12-9　【例 12.6】试验资料方差分析表

变异来源	SS	df	MS	F	$F_{0.05(1,4)}$
处　理	1.6748	1	1.6748	1.60^{ns}	7.71
误　差	4.1727	4	1.0432		
总变异	17.72	5			

F 检验结果表明，在对照饲料基础上添加尿素对提高奶牛产奶量效果不显著。

（二）t 检验法　2×3 交叉试验资料分析的 t 检验公式与 2×2 交叉试验资料分析的 t 检验相同。

此例　$r=s=3$,　$T_1=3.06$,　$T_2=-0.11$,　$\bar{d}_1=1.0200$,　$\bar{d}_2=-0.0367$

$$S_{\bar{d}_1-\bar{d}_2}=\sqrt{\frac{[(-0.09)^2+0.70^2+2.45^2-\frac{3.06^2}{3}]+[0.39^2+(-0.76)^2+0.26^2-\frac{(-0.11)^2}{3}]}{(3-1)+(3-1)}\left(\frac{1}{3}+\frac{1}{3}\right)}$$

$$=0.8339$$

$$t=\frac{\bar{d}_1-\bar{d}_2}{S_{\bar{d}_1-\bar{d}_2}}=\frac{1.02-(-0.0367)}{0.8339}=1.2672$$

由 $df=(r-1)+(s-1)=(3-1)+(3-1)=4$ 查临界 t 值，得：$t_{0.05(4)}=2.776$，因为 $t=1.2672<t_{0.05(4)}$，$P>0.05$，表明在对照饲料上添加尿素与否，奶牛产奶量差异不显著。检验结果与方差分析法相同。

三、交叉设计的优缺点及注意事项

（一）交叉设计的优缺点

1、主要优点　交叉设计可以消除个体间及试验时期间的差异对试验结果的影响，进一步突出处理效应，提高了试验的精确性。因此，交叉设计特别适用于个体差异较大的动物试验，如大动物和兽医学试验等。此外，交叉试验结果的分析较为简便。

2、主要缺点　与拉丁方设计相比，交叉设计不能得到关于个体差异和试验期差异大小的信息；若与有重复的多因素试验相比，还不能得到因素之间交互作用的信息。因此，交叉设计适用范围有一定的局限性。

（二）应用交叉设计须注意的问题

1、处理因素、时期、个体间不存在交互作用　如果交叉试验中处理因素、时期、个体有交互作用，这些交互作用效应就会归入误差项中，使误差估计值增大，从而降低试验的精确性。

2、要注意试验是否有处理残效　在交叉试验中，处理轮流更换，如果前一种处理有效应残存，则观测值的线性模型条件就不能成立。为解决这个问题，可设置适当的预试期和间歇期。对于残效不能消失的处理，例如带有破坏性且不能恢复的试验，则不宜采用交叉设计。

3、采用 **Lucas** 提出的方差分析法分析 2×2、2×3 交叉试验资料时要求各试验组动物的头数相等　如在【例 12.6】中，第一组按 A_1—A_2—A_1 的顺序给予饲料，第二组按 A_2—A_1—A_2 顺序给予饲料，每头奶牛被分配到哪一组，是随机确定的，但两个组的奶牛头数必须相等。只有这样才能通过 Σd_{1j}-Σd_{2j} 而使试验期的效应相互抵消。**采用明道绪提出的 *t* 检验法分析 2×2、2×3 交叉试验资料时，不要求两组试验个体数相等**。因而 *t* 检验法应用范围更广，且计算步骤也较为简明。

*第八节　正交设计

在动物试验研究中，对于单因素或两因素试验，因其因素少，试验的设计、实施与分析都比较简单。但在实际工作中，常常需要同时考察 3 个或 3 个以上的试验因素，若进行全面试验，则试验的规模将很大，往往因试验条件的限制而难于实施。正交设计就是安排多因素试验、寻求最优水平组合的一种高效率试验设计方法。

一、正交设计的概念及原理

（一）正交设计的基本概念　正交设计是利用正交表来安排与分析多因素试验的一种设计方法。它利用从试验的全部水平组合中，挑选部分有代表性的水平组合进行试验，通过对这部分试验结果的分析了解全面试验的情况，找出最优的水平组合。

例如，影响某品种鸡的生产性能有 3 个因素：A 因素是饲料配方，分 A_1、A_2、A_3 3 个水平；B 因素是光照，分 B_1、B_2、B_3 3 个水平；C 因素是温度，分 C_1、C_2、C_3 3 个水平。

这是一个 3 因素 3 水平的试验，各因素的水平之间全部可能的组合有 27 种。如果试验方案包含各因素的全部水平组合，即进行全面试验，可以分析各因素的效应，交互作用，也可选出最优水平组合。这是全面试验的优点。但全面试验包含的水平组合数较多，工作量大，由于受试验场地、试验动物、经费等限制而难于实施。若试验的目的主要是寻求最优水平组合，则可利用正交设计来安排试验。正交设计的基本特点是：用部分试验来代替全面试验，通过对部分试验结果的分析，了解全面试验的情况。正因为正交试验是用部分试验来代替全面试验，它不可能像全面试验那样对各因素效应、交互作用一一分析；当交互作用存在时，有可能出现交互作用的混杂。虽然正交设计有上述不足，但它能通过部分试验找到最优水平组合，因而很受实际工作者青睐。

如对于上述 3 因素 3 水平试验，若不考虑交互作用，可利用正交表 $L_9(3^4)$安排，试验方案仅包含 9 个水平组合，就能反映试验方案包含 27 个水平组合的全面试验的情况，找出最佳的生产条件。

（二）正交设计的基本原理 在试验安排中，每个因素在研究的范围内选几个水平，就好比在选优区内打上网格，如果网上的每个点都做试验，就是全面试验。如上例中，3 个因素的选优区可以用一个立方体表示(图 12-2)，3 个因素各取 3 个水平，把立方体划分成 27 个格点，反映在图 12-2 上就是立方体内的 27 个“.”。若 27 个网格点都试验，就是全面试验，其试验方案如表 12-20 所示。

表 12-20　3 因素 3 水平全面试验方案

		C_1	C_2	C_3
A_1	B_1	$A_1B_1C_1$	$A_1B_1C_2$	$A_1B_1C_3$
	B_2	$A_1B_2C_1$	$A_1B_2C_2$	$A_1B_2C_3$
	B_3	$A_1B_3C_1$	$A_1B_3C_2$	$A_1B_3C_3$
A_2	B_1	$A_2B_1C_1$	$A_2B_1C_2$	$A_2B_1C_3$
	B_2	$A_2B_2C_1$	$A_2B_2C_2$	$A_2B_2C_3$
	B_3	$A_2B_3C_1$	$A_2B_3C_2$	$A_2B_3C_3$
A_3	B_1	$A_3B_1C_1$	$A_3B_1C_2$	$A_3B_1C_3$
	B_2	$A_3B_2C_1$	$A_3B_2C_2$	$A_3B_2C_3$
	B_3	$A_3B_3C_1$	$A_3B_3C_2$	$A_3B_3C_3$

图 12-2　3 因素 3 水平试验的均衡分散立体图

3 因素 3 水平的全面试验水平组合数为 3^3=27，4 因素 3 水平的全面试验水平组合数为 3^4=81，5 因素 3 水平的全面试验水平组合数为 3^5=243，这在动物试验中是不可能做到的。正交设计就是从选优区全面试验点（水平组合）中挑选出有代表性的部分试验点（水平组合）来进行试验。图 12-2 中标有试验号的九个“⊙”，就是利用正交表 $L_9(3^4)$从 27 个试验点中挑选出来的 9 个试验点。即：

(1)$A_1B_1C_1$　　(2)$A_2B_1C_2$　　(3)$A_3B_1C_3$

(4)$A_1B_2C_2$　　(5)$A_2B_2C_3$　　(6)$A_3B_2C_1$

(7)$A_1B_3C_3$　　(8)$A_2B_3C_1$　　(9)$A_3B_3C_2$

上述选择，保证了 A 因素的每个水平与 B 因素、C 因素的各个水平在试验中各搭配一次。对于 A、B、C 3 个因素来说，是在 27 个全面试验点中选择 9 个试验点，仅是全面试验的三分之一。从图 12-2 中可以看到，9 个试验点在选优区中分布是均衡的，在立方体的每个平面上，都恰是 3 个试验点；在立方体的每条线上也恰有一个试验点。9 个试验点均衡地分布于整个立方体内，有很强的代表性，能够比较全面地反映选优区内的基本情况。

二、正交表及其特性

（一）正交表　由于正交设计安排试验和分析试验结果都要用正交表，因此，我们先对正交表作一介绍。表 12-21 是一张正交表，记号为 $L_8(2^7)$，其中“L”代表正交表；L 右下角的数字“8”表示有 8 行，用这张正交表安排试验包含 8 个处理(水平组合)；括号内的底数“2”表示因素的水平数，括号内 2 的指数“7”表示有 7 列，用这张正交表最多可以安排 7 个因素。

表 12-21　$L_8(2^7)$正交表

试验号	列号						
	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

常用的正交表已由数学工作者制定出来，供进行正交设计时选用。2 水平正交表有 $L_4(2^3)$、$L_{16}(2^{15})$；3 水平正交表有 $L_9(3^4)$、$L_{27}(2^{13})$……等（详见附表 14 及有关参考书）。

（二）正交表的特性　任何一张正交表都有如下两个特性：

1、*任一列中，不同数字出现的次数相等*　例如 $L_8(2^7)$中不同数字只有 1 和 2，它们各出现 4 次；$L_9(3^4)$中不同数字有 1、2 和 3，它们各出现 3 次。

2、*任两列中，同一横行所组成的数字对出现的次数相等*　例如 $L_8(2^7)$中(1, 1), (1, 2), (2, 1), (2, 2)各出现两次；$L_9(3^4)$中(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3)各出现 1 次。即每个因素的一个水平与另一因素的各个水平互碰次数相等，表明任意两列各个数字之间的搭配是均匀的。

根据以上两个特性，我们用正交表安排的试验，具有均衡分散和整齐可比的特点。所谓均衡分散，是指用正交表挑选出来的各因素水平组合在全部水平组合中的分布是均匀的。由图 12-2 可以看出，在立方体中，任一平面内都包含 3 个“⊙”，任一直线上都包含 1 个“⊙”，因此，这些点代表性强，能够较好地反映全面试验的情况。整齐可比是指每一个因素的各水

平间具有可比性。因为正交表中每一因素的任一水平下都均衡地包含着另外因素的各个水平，当比较某因素不同水平时，其它因素的效应都彼此抵消。如在 A、B、C 3 个因素中，A 因素的 3 个水平 A_1、A_2、A_3 条件下各有 B、C 的 3 个不同水平，即：

	B_1C_1		B_1C_2		B_1C_3
A_1	B_2C_2	A_2	B_2C_3	A_3	B_2C_1
	B_3C_3		B_3C_1		B_3C_2

在这 9 个水平组合中，A 因素各水平下包括了 B、C 因素的 3 个水平，虽然搭配方式不同，但 B、C 皆处于同等地位，当比较 A 因素不同水平时，B 因素不同水平的效应相互抵消，C 因素不同水平的效应也相互抵消。所以 A 因素 3 个水平间具有可比性。同样，B、C 因素 3 个水平间亦具有可比性。

（三）正交表的类别

1、*相同水平正交表*　各列中出现的最大数字相同的正交表称为相同水平正交表。如 $L_4(2^3)$、$L_8(2^7)$、$L_{12}(2^{11})$等各列中最大数字为 2，称为两水平正交表；$L_9(3^4)$、$L_{27}(3^{13})$等各列中最大数字为 3，称为 3 水平正交表。

2、*混合水平正交表*　各列中出现的最大数字不完全相同的正交表称为混合水平正交表。如 $L_8(4\times2^4)$表中有一列最大数字为 4，有 4 列最大数字为 2。也就是说该表可以安排一个 4 水平因素和 4 个 2 水平因素。再如 $L_{16}(4^4\times2^3)$，$L_{16}(4\times12^{12})$等都混合水平正交表。

三、正交设计方法

【例 12.7】　在进行矿物质元素对架子猪补饲试验中，考察补饲配方、用量、食盐 3 个因素，每个因素都有 3 个水平。试安排一个正交试验方案。

正交设计一般有以下几个步骤：

（一）确定因素和水平　影响试验结果的因素很多，我们不可能把所有影响因素通过一次试验都予以研究，只能根据以往的经验，挑选和确定若干对试验指标影响最大、有较大经济意义而又了解不够清楚的因素来研究。同时还应根据实际经验和专业知识，定出各因素适宜的水平，列出因素水平表。**【例 12.7】**的因素水平表如表 12-22 所示。

表 12-22　架子猪补饲试验因素水平表

水　平	因　素		
	矿物质元素补饲配方(A)	用　量(g)(B)	食　盐(g)(C)
1	配方 I(A_1)	15(B_1)	0(C_1)
2	配方 II(A_2)	25(B_2)	4(C_2)
3	配方 III(A_3)	20(B_3)	8(C_3)

（二）选用合适的正交表　确定了因素及其水平后，根据因素、水平及需要考察的交互作用的多少来选择合适的正交表。选用正交表的原则是：既要能安排下试验的全部因素，又要使部分水平组合数（处理数）尽可能地少。一般情况下，试验因素的水平数应恰好等于正交表记号中括号内的底数；因素的个数（包括交互作用）应不大于正交表记号中括号内的指数；各因素及交互作用的自由度之和要小于所选正交表的总自由度，以便估计试验误差。

若各因素及交互作用的自由度之和等于所选正交表总自由度，则可采用有重复正交试验来估计试验误差。

此例有 3 个 3 水平因素，若不考察交互作用，则各因素自由度之和为因素数个数×(水平数—1)=3(3-1)=6，小于 $L_9(3^4)$总自由度 9-1=8，故可以选用 $L_9(3^4)$；若要考察交互作用，则应选用 $L_{27}(3^{13})$，此时所安排的试验方案实际上是全面试验方案。

（三）表头设计 正交表选好后，就可以进行表头设计。所谓表头设计，就是把挑选出的因素和要考察的交互作用分别排入正交表的表头适当的列上。在不考察交互作用时，各因素可随机安排在各列上；若考察交互作用，就应按该正交表的交互作用列表安排各因素与交互作用。此例不考察交互作用，可将矿物质元素补饲配方(A)、用量(B)和食盐(C)依次安排在 $L_9(3^4)$的第 1、2、3 列上，第 4 列为空列，见表 12-23。

表 12-23　表头设计

列　　号	1	2	3	4
因　　素	A	B	C	空

（四）列出试验方案 把正交表中安排各因素的每个列(不包含欲考察的交互作用列)中的每个数字依次换成该因素的实际水平，就得到一个正交试验方案。表 12-24 就是[例 12.4]的正交试验方案。

根据表 12-24，1 号试验处理是 $A_1B_1C_1$，即配方 I、用量 15g、食盐为 0；2 号试验处理是 $A_1B_2C_2$，即配方 II、用量 25g、食盐为 4g，…；9 号试验处理为 $A_3B_3C_2$，即配方 III、用量 20g、食盐 4g。

表 12-24　正交试验方案

试　验　号	因　　素		
	A	B	C
	1	2	3
1	1(配方 I)	1(15)	1(0)
2	1(配方 I)	2(25)	2(4)
3	1(配方 I)	3(20)	3(8)
4	2(配方 II)	1(15)	2(4)
5	2(配方 II)	2(25)	3(8)
6	2(配方 II)	3(20)	1(0)
7	3(配方 III)	1(15)	3(8)
8	3(配方 III)	2(25)	1(0)
9	3(配方 III)	3(20)	2(4)

四、正交试验结果的统计分析

根据各号试验处理是单独观测值还是有重复观测值，正交试验可分为单独观测值正交试验和有重复观测值正交试验两种。若各号试验处理都只有一个观测值，则称之为单独观测值正交试验；若各号试验处理都有两个或两个以上观测值，则称之为有重复观测值正交试验。

下面分别介绍单独观测值和有重复观测正交试验结果的方差分析。

（一）单独观测值正交试验结果的方差分析 对例【12.7】用 $L_9(3^4)$安排试验方案后，各号试验只进行一次，试验结果(增重)列于表 12-17。试对其进行方差分析。

该次试验的 9 个观测值总变异由 A 因素、B 因素、C 因素及误差变异四部分组成，因而进行方差分析时平方和与自由度的划分式为：

$$SS_T = SS_A+SS_B+SS_C+SS_e$$
$$df_T = df_A+df_B+df_C+df_e \qquad (12\text{-}6)$$

用 n 表示试验(处理)号数；a、b、c 表示 A、B、C 因素各水平重复数；k_a、k_b、k_c 表示 A、B、C 因素的水平数。本例，n=9、a=b=c=3、 k_a=k_b=k_c=3。

表 12-17　正交试验结果计算表

试验号	因素 A (1)	B (2)	C (3)	增重(kg)
1	1	1	1	63.4 (y_1)
2	1	2	2	68.9 (y_2)
3	1	3	3	64.9 (y_3)
4	2	1	2	64.3 (y_4)
5	2	2	3	70.2 (y_5)
6	2	3	1	65.8 (y_6)
7	3	1	3	71.4 (y_7)
8	3	2	1	69.5 (y_8)
9	3	3	2	73.7 (y_9)
T_1	197.2	199.1	198.7	612.1(T)
T_2	200.3	208.6	206.9	
T_3	214.6	204.4	206.5	
$\bar{x}_1$	65.7333	66.3667	66.2333	
$\bar{x}_2$	66.7667	69.5333	68.9667	
$\bar{x}_3$	71.5333	68.1333	68.8333	

表 12-17 中，T_i 为各因素同一水平试验指标（增重）之和。如 A 因素第 1 水平 T_1=y_1+y_2+y_3=63.4+68.9+64.9=197.2，A 因素第 2 水平 T_2=y_4+y_5+y_6=64.3+70.2+65.8=200.3，A 因素第 3 水平 T_3=y_7+y_8+y_9=71.4+69.5+73.7=214.6；B 因素第 1 水平 T_1=y_1+y_4+y_7=63.4+64.3+71.4=199.1，……，B 因素第 3 水平 T_3=y_3+y_6+y_9=64.9+65.8+73.7=204.4。同理可求得 C 因素各水平试验指标之和。

$\bar{x}$ 为各因素同一水平试验指标的平均数。如 A 因素第 1 水平 $\bar{x}_1$ =197.2/3=65.7333，A 因素第 2 水平 $\bar{x}_2$ =200.3/3=66.7667，A 因素第 3 水平 $\bar{x}_3$ =214.6/3=71.5333。同理可求得 B、C 因素各水平试验指标的平均数。

1、计算各项平方和与自由度

矫正数　　$C=T^2/n=612.1^2/9=41629.6011$

总平方和　　$SS_T=\Sigma y^2-C=63.4^2+68.9^2+\cdots+73.7^2-41629.6011=101.2489$

A 因素平方和　　$SS_A=\Sigma T^2_A/a-C=(197.2^2+200.3^2+214.6^2)/3-41629.6011=57.4289$

B 因素平方和　　$SS_B=\Sigma T^2_B/b-C=(199.1^2+208.6^2+204.4^2)/3-41629.6011=15.1089$

C 因素平方和　　$SS_C=\Sigma T^2_C/c-C=(198.7^2+206.9^2+206.5^2)/3-41629.6011=14.2489$

误差平方和　　$SS_e=SS_T-SS_A-SS_B-SS_C=101.2489-57.4289-15.1089-14.2489=14.4622$

总自由度　　$df_T=n-1=9-1=8$

A 因素自由度　$df_A=k_a-1=3-1=2$

B 因素自由度　$df_B=k_b-1=3-1=2$

C 因素自由度　$df_C=k_c-1=3-1=2$

误差自由度　$df_e=df_T-df_A-df_B-df_C=8-2-2-2=2$

2、列出方差分析表，进行 *F* 检验

表 12-26　方差分析表

变异来源	*SS*	*df*	*MS*	*F*	$F_{0.05(2,2)}$
配方(A)	57.4289	2	28.71	3.97^{ns}	19.00
用量(B)	15.1089	2	7.55	1.05^{ns}	
食盐(C)	14.2489	2	7.12	<1	
误差	14.4622	2	7.23		
总变异	101.25	8			

F 检验结果表明，三个因素对增重的影响都不显著。究其原因可能是本例试验误差大且误差自由度小(仅为 2)，使检验的灵敏度低，从而掩盖了考察因素的显著性。由于各因素对增重影响都不显著，不必再进行各因素水平间的多重比较。此时，可直观地从表 12-17 中选择平均数大的水平 A_3、B_3、C_2 组合成最优水平组合 $A_3B_3C_2$。

上述无重复正交试验结果的方差分析，其误差是由“空列”来估计的。然而“空列”并不空，实际上是被未考察的交互作用所占据。这种误差既包含试验误差，也包含交互作用，称为模型误差。若交互作用不存在，用模型误差估计试验误差是可行的；若因素间存在交互作用，则模型误差会夸大试验误差，有可能掩盖考察因素的显著性。这时，试验误差应通过重复试验值来估计。所以，进行正交试验最好能有二次以上的重复。正交试验的重复，可采用完全随机或随机单位组设计。

（二）有重复观测值正交试验结果的方差分析　假定【例 12.7】试验重复了两次，且重复采用随机单位组设计，试验结果列于表 12-27。试对其进行方差分析。

用 n 表示试验(处理)号数，r 表示试验处理的重复数。a、b、c、k_a、k_b、k_c 的意义同上。此例 n=9、r=2、a=b=c=3、$k_a=k_b=d_c=3$

表 12-27　有重复观测值正交试验结果计算表

试验号	因素				增重 (kg)		
	A (1)	B (2)	C (3)	空 (4)	单位组 I	单位组 II	T_t
1	1	1	1	1	63.4	67.4	130.8
2	1	2	2	2	68.9	87.2	156.1
3	1	3	3	3	64.9	66.3	131.2
4	2	1	2	3	64.3	86.3	150.6
5	2	2	3	1	70.2	88.5	158.7
6	2	3	1	2	65.8	66.6	132.4
7	3	1	3	2	71.4	89.0	160.4
8	3	2	1	3	69.5	91.2	160.7
9	3	3	2	1	73.7	92.8	166.5
T_1	418.1	441.8	423.9	456	612.1	735.3	1347.4(T)
T_2	441.7	475.5	473.2	448.9			
T_3	487.6	430.1	450.3	442.5			
$\bar{x}_1$	69.68	73.63	70.65	76.00			
$\bar{x}_2$	73.62	79.25	78.87	74.82			
$\bar{x}_3$	81.26	71.68	75.05	73.75			

对于有重复、且重复采用随机单位组设计的正交试验，总变异可以划分为处理间、单位组间和误差变异三部分，而处理间变异可进一步划分为 A 因素、B 因素、C 因素与模型误差变异四部分。此时，平方和与自由度划分式为：

$$SS_T=SS_t+SS_r+SS_{e2}$$

$$df_T=df_t+df_r+df_{e2}$$

而

$$SS_t=SS_A+SS_B+SS_C+SS_{e1}$$

$$df_t=df_A+df_B+df_C+df_{e1}$$

于是

$$SS_T=SS_A+SS_B+SS_C+SS_r+SS_{e1}+SS_{e2} \quad (12\text{-}7)$$

$$df_T=df_A+df_B+df_C+df_r+df_{e1}+df_{e2}$$

式中：SS_r 为单位组间平方和；SS_{e1} 为模型误差平方和；SS_{e2} 为试验误差平方和；SS_t 为处理间平方和；df_r、df_{e1}、df_{e2} 、df_t 为相应自由度。

注意，对于重复采用完全随机设计的正交试验，在平方和与自由度划分式中无 SS_r、df_r 项。

1、计算各项平方和与自由度

矫正数　$C=T^2/rn=1347.4^2/2\times9=100860.3756$

总平方和　$SS_T=\Sigma y^2-C=63.4^2+68.9^2+\cdots+92.8^2-100860.3756=1978.5444$

单位组间平方和　$SS_r=\Sigma T^2_r/n-C=(612.1^2+735.3^2)/9-100860.3756=843.2355$

处理间平方和　$SS_t=\Sigma T^2_t/r-C=(130.8^2+156.1^2+\cdots+166.5^2)/2-100860.3756=819.6244$

A 因素平方和　$SS_A=\Sigma T^2_A/ar-C=(418.1^2+441.7^2+487.6^2)/3\times2-100860.3756=416.3344$

B 因素平方和　$SS_B=\Sigma T^2_B/br-C=(411.8^2+475.5^2+430.1^2)/3\times2-100860.3756=185.2077$

C 因素平方和　$SS_C=\Sigma T^2_C/cr-C=(423.9^2+473.2^2+450.3^2)/3\times2-100860.3756=202.8811$

模型误差平方和　$SS_{e1}=SS_t-SS_A-SS_B-SS_C=819.6244-416.3344-185.2077-202.8811=15.2012$

试验误差平方和　$SS_{e2}=SS_T-SS_r-SS_t=1978.5444-843.2355-819.6244=315.6845$

总自由度　$df_T=rn-1=2\times9-1=17$

单位组自由度　$df_r=r-1=2-1=1$

处理自由度　$df_t=n-1=9-1=8$

A 因素自由度　$df_A=a-1=3-1=2$

B 因素自由度　$df_B=b-1=3-1=2$

C 因素自由度　$df_C=c-1=3-1=2$

模型误差自由度　$df_{e1}=df_t-df_A-df_B-df_C=8-2-2-2-2=2$

试验误差自由度　$df_{e2}=df_T-df_t=17-1-8=8$

2、列出方差分析表，进行 F 检验

表 12-8　有重复观测值正交试验结果方差分析表

变异来源	*SS*	*df*	*MS*	*F*	$F_{0.05}$	$F_{0.01}$
A	416.3344	2	208.17	6.29*	4.10	7.55
B	185.2077	2	92.60	2.80	4.10	7.55
C	202.8811	2	101.44	3.07	4.10	7.55
单位组	843.2355	1	843.24	25.48**	4.96	10.01

误差(e1)	15.2012	2	7.60
误差(e2)	315.6845	8	39.46
合并误差	330.8857	10	33.09
总　　的	1978.5444	17	

首先检验 MS_{e1} 与 MS_{e2} 差异的显著性，若经 F 检验不显著，则可将其平方和与自由度分别合并，计算出合并的误差均方，进行 F 检验与多重比较，以提高分析的精度；若 F 检验显著，说明存在交互作用，二者不能合并，此时只能以 MS_{e2} 进行 F 检验与多重比较。本例 $MS_{e1}/MS_{e2}<1$，MS_{e1} 与 MS_{e2} 差异不显著，故将误差平方和与自由度分别合并计算出合并的误差均方 MSe，即 $MSe=(SS_{e1}+SS_{e2})/(df_{e1}+df_{e2})=(15.2012+315.6845)/(2+8)=33.09$，并用合并的误差均方 MSe 进行 F 检验与多重比较。

F 检验结果表明，矿物质元素配方对架子猪增得有显著影响，另外两个因素作用不显著；二个单位组间差异极显著。

3、 *A* 因素各水平平均数的多重比较

表 12-9　*A* 因素各水平平均数多重比较表(*SSR* 法)　　单位：kg

A 因素	平均数 $\bar{x}_i$	$\bar{x}_i$ -69.68	$\bar{x}_i$ -73.62
A_3	81.26	11.58**	7.64*
A_2	73.62	3.94	
A_1	69.68		

因为，$S_{\bar{x}}=\sqrt{MS_e/ar}=\sqrt{33.09/(3\times2)}=2.35$

由 df_e=10 和 k=2, 3, 查得 *SSR* 值并计算出 *LSR* 值列于表 12-30。

表 12-30　*SSR* 值与 *LSR* 值表

df_e	k	$SSR_{0.05}$	$SSR_{0.01}$	$LSR_{0.05}$	$LSR_{0.01}$
10	2	3.15	4.48	7.40	10.53
	3	3.30	4.73	7.76	11.12

多重比较结果表明：A 因素 A_3 水平的平均数显著或极显著地高于 A_2、A_1；A_2 与 A_1 间差异不显著。

此例因模型误差不显著，可以认为因素间不存在显著的交互作用。可由 A、B、C 因素的最优水平组合成最优水平组合。A 因素的最优水平为 A_3；因为 B、C 因素水平间差异均不显著，故可任选一水平。如 B、C 因素选择使增重达较高水平的 B_2 及 C_2，则得最优水平组合为 $A_3B_2C_2$，即配方 III、用量 25 克、食盐 4 克。

若模型误差显著，表明因素间交互作用显著，则应进一步试验，以分析因素间的交互作用。

五、因素间有交互作用的正交设计与分析

在实际研究中，有时试验因素之间存在交互作用。对于既考察因素主效应又考察因素间交互作用的正交设计，除表头设计和结果分析与前面介绍略有不同外，其它基本相

同。

【例 12.8】 某一种抗菌素的发酵培养基由 *A*、*B*、*C* 3 种成分组成，各有两个水平，除考察 *A*、*B*、*C* 三个因素的主效外，还考察 *A* 与 *B*、*B* 与 *C* 的交互作用。试安排一个正交试验方案并进行结果分析。

（一）选用正交表，作表头设计 由于本试验有 3 个两水平的因素和两个交互作用需要考察，各项自由度之和为：3×(2-1)+2×(2-1)×(2-1)=5，因此可选用 $L_8(2^7)$来安排试验方案。

正交表 $L_8(2^7)$中有基本列和交互列之分，基本列就是各因素所占的列，交互列则为两因素交互作用所占的列。可利用 $L_8(2^7)$二列间交互作用列表(见表 12-31)来安排各因素和交互作用。

表 12-31　$L_8(2^7)$二列间交互作用列表

列号	1	2	3	4	5	6	7
1	(1)	3	2	5	4	7	6
2		(2)	1	6	7	4	5
3			(3)	7	6	5	4
4				(4)	1	2	3
5					(5)	3	2
6						(6)	1

如果将 A 因素放在第 1 列，B 因素放在第 2 列，查表 12-31 可知，第 1 列与第 2 列的交互作用列是第 3 列，于是将 A 与 B 的交互作用 A×B 放在第 3 列。这样第 3 列不能再安排其它因素，以免出现“混杂”。然后将 C 放在第 4 列，查表 12-31 可知，B×C 应放在第 6 列，余下列为空列，如此可得表头设计，见表 12-32。

表 12-32　表头设计

列号	1	2	3	4	5	6	7
因素	A	B	A×B	C	空	B×C	空

（二）列出试验方案 根据表头设计，将 A、B、C 各列对应的数字“1”、“2”换成各因素的具体水平，得出试验方案列于表 12-33。

表 12-33　正交试验方案

试 验 号	因　素		
	1(A)	2(B)	3(C)
1	$1(A_1)$	$1(B_1)$	$1(C_1)$
2	$1(A_1)$	$1(B_1)$	$2(C_2)$
3	$1(A_1)$	$2(B_2)$	$1(C_1)$
4	$1(A_1)$	$2(B_2)$	$2(C_2)$
5	$2(A_2)$	$1(B_1)$	$1(C_1)$
6	$2(A_2)$	$1(B_1)$	$2(C_2)$
7	$2(A_2)$	$2(B_2)$	$1(C_1)$

8	2(A_2)	2(B_2)	2(C_2)

（三）结果分析 按表 12-33 所列的试验方案进行试验，其结果见表 12-34。

表中 T_i、$\bar{x}_i$ 计算方法同前。此例为单独观测值正交试验，总变异划分为 A 因素、B 因素、C 因素、A×B、B×C、与误差变异 5 部分，平方和与自由度划分式为：

$$SS_T=SS_A+SS_B+SS_C+SS_{A\times B}+SS_{B\times C}+SS_e$$

$$df_T=df_A+df_B+df_C+df_{A\times B}+df_{B\times C}+df_e \qquad (12\text{-}8)$$

1、计算各项平方和与自由度

矫正数 $C=T^2/n=665^2/8=55278.1250$

总平方和 $SS_T=\Sigma y^2-C=55^2+38^2+\cdots+61^2-55278.1250=6742.8750$

A 因素平方和 $SS_A=\Sigma T_A^2/a-C=(279^2+386^2)/4-55278.1250=1431.1250$

B 因素平方和 $SS_B=\Sigma T_B^2/b-C=(339^2+326^2)/4-55278.1250=21.1250$

C 因素平方和 $SS_C=\Sigma T_C^2/c-C=(353^2+312^2)/4-55278.1250=210.1250$

A×B 平方和 $SS_{A\times B}=\Sigma T_{A\times B}^2/4-C=(233^2+432^2)/4-55278.1250=4950.1250$

B×C 平方和 $SS_{B\times C}=\Sigma T_{B\times C}^2/4-C=(327^2+338^2)/4-55278.1250=15.1250$

误差平方和 $SS_e=SS_T-SS_A-SS_B-SS_{A\times B}-SS_{B\times C}=6742.8750-1431.1250-21.1250$
$-210.1250-4950.1250-15.1250=115.2500$

总自由度 $df_T=n-1=8-1=7$

各因素自由度 $df_A=df_B=df_C=2-1=1$

交互作用自由度 $df_{A\times B}=df_{B\times C}=(2-1)(2-1)=1$

误差自由度 $df_e=df_T-df_A-df_C-df_{A\times B}-df_{B\times C}=7-1-1-1-1-1=2$

表 12-34 有交互作用的正交试验结果计算表

试验号	因素					试验结果(%)*
	A	B	A×B	C	B×C	
1	1	1	1	1	1	55(y_1)
2	1	1	1	2	2	38(y_2)
3	1	2	2	1	2	97(y_3)
4	1	2	2	2	1	89(y_4)
5	2	1	2	1	1	122(y_5)
6	2	1	2	2	2	124(y_6)
7	2	2	1	1	2	79(y_7)
8	2	2	1	2	1	61(y_8)
T_1	279	339	233	353	327	665(T)
T_2	386	326	432	312	338	
$\bar{x}_1$	69.75	84.75	58.25	88.25	81.75	
$\bar{x}_2$	96.50	81.50	108.00	78.00	84.50	

*试验结果以对照为 100 计

2、列出方差分析表，进行 *F* 检验

表 12-35 方差分析表

变异来源	*SS*	*df*	*MS*	*F*	$F_{0.05(1,2)}$	$F_{0.01(1,2)}$
A	1431.1250	1	1431.1250	24.84*	18.51	98.49
B	21.1250	1	21.1250	<1		
C	210.1250	1	210.1250	3.65		

A×*B*	4950.1250	1	4950.1250	85.90*
B×*C*	15.1250	1	12.1250	<1
误差	115.1250	2	57.6250	
总的	6742.8750	7		

F 检验结果表明：A 因素和交互作用 A×B 显著，B、C 因素及 B×C 交互作用不显著。因交互作用 A×B 显著，应对 A 与 B 的水平组合进行多重比较，以选出 A 与 B 的最优水平组合。

3、A 与 B 各水平组合的多重比较

先计算出 A 与 B 各水平组合的平均数：

A_1B_1 水平组合的平均数 $\bar{x}_{11}$=(55+38)/2=46.50

A_1B_2 水平组合的平均数 $\bar{x}_{12}$=(97+89)/2=93.00

A_2B_1 水平组合的平均数 $\bar{x}_{21}$=(122+124)/2=123.00

A_2B_2 水平组合的平均数 $\bar{x}_{22}$=(79+61)/2=70.00

列出 A、B 因素各水平组合平均数多重比较表，见表 12-36。

表 12-36　A、B 因素各水平组合平均数多重比较表(q 法)

水平组合	平均数	$\bar{x}_{ij}$ -46.5	$\bar{x}_{ij}$ -70	$\bar{x}_{ij}$ -93
A_2B_1	123.00	76.5*	53*	30
A_1B_2	93.00	46.5*	23	
A_2B_2	70.00	23.5		
A_1B_1	46.50			

因为，$S_{\bar{x}}=\sqrt{MS_e/2}=\sqrt{57.625/2}=5.37$，由 df_e=2 与 k=2, 3, 4, 查临界 q 值，并计算出 *LSR* 值，见表 12-37。

表 12-37　*q* 值与 *LSR* 值表

dfe	*k*	$q_{0.05}$	$q_{0.01}$	$LSR_{0.05}$	$LSR_{0.01}$
2	2	6.09	14.0	32.70	75.18
	3	8.28	19.0	44.46	102.03
	4	9.80	22.3	52.63	119.75

多重比较结果表明，A_2B_1 显著优于 A_2B_2，A_1B_1；A_1B_2 显著优于 A_1B_1，其余差异不显著。最优水平组合为 A_2B_1。

从以上分析可知，A 因素取 A_2，B 因素取 B_1，若 C 因素取 C_1，则本次试验结果的最优水平组合为 $A_2B_1C_1$。

注意，此例因 df_e=2，F 检验与多重比较的灵敏度低。为了提高检验的灵敏度，可将 F<1 的 SS_B、df_B，$SS_{B\times C}$、$df_{B\times C}$ 合并到 SS_e、df_e 中，得合并的误差均方，再用合并误差均方进行 F 检验与多重比较。这一工作留给读者完成。

第九节　调查设计

在科学研究中，除了进行控制试验外，有时也要进行调查研究。调查研究是对已有的事实通过各种方式进行了解，然后用统计的方法对所得数据进行分析，从而找出其中的规律性。例如，了解畜禽品种及水产资源状况；探索和分析对某种疾病有效的防治规律、措施以及新的检验手段和方法等。由于现场调查立足于生产实际，所以它是研究和解决实际问题的一种重要研究方法。同时，控制试验的研究课题，往往是在调查研究的基础上确定的；试验研究的成果，又必须在其推广应用后经调查得以验证。

为了使调查研究工作有目的、有计划、有步骤地顺利开展，必须事先拟定一个详细的调查计划。调查计划应包括以下几个内容：

（一）调查研究的目的 任何一项调查研究都要有明确的目的，即通过调查了解什么问题，解决什么问题。例如，家畜健康状况的调查的目的是评定家畜健康水平；畜禽品种资源调查的目的是了解畜禽品种的数量、分布与品种特征特性等情况。同时，调查研究的目的还应该突出重点，一次调查应针对主要问题收集必要的数据，深入分析，为主要问题的解决提出相应的措施和办法。

（二）调查的对象与范围 根据调查的目的，确定调查的对象、地区和范围，划清调查总体的同质范围、时间范围和地区范围。例如，四川省家禽品种资源调查，调查地区为四川省，调查总体和对象为全省各市、县的家禽，调查时间从 2000 年 1 月到 2000 年 12 月。

（三）调查的项目 调查项目的确定要紧紧围绕调查目的。调查项目确定的正确与否直接关系到调查的质量。因此，项目应尽量齐全，重要的项目不能漏掉；项目内容要具体、明确，不能模棱两可。应按不同的指标顺序以表格形式列示出来，以达到顺利完成搜集资料的目的。例如，家禽品种资源调查项目有：种类(鸡、鸭、鹅等)、品种(柴鸡、来航、白洛克等)，数量、体重、产蛋性能等项目。

调查项目有一般项目和重点项目之分。一般项目主要是指调查对象的一般情况，用于区分和查找，如畜主姓名、住址及编号等。重点项目是调查的核心内容，如品种资源调查中的品种、数量及生产性能等。

调查表的形式分为一览表和卡片，当调查的指标较少时多采用一览表的形式，它可以填入许多调查动物情况。若调查的内容多而复杂时可采用卡片的形式，一张卡片只填一个对象，以便汇总和整理，或输入计算机。

（四）样本含量 在抽样调查研究时，样本含量的大小关系到调查结果的精确性。样本含量太大，需耗费较多的人力、物力及资金；样本含量太小，增大了偶然性，使抽样误差大，影响调查结果的精确性。确定样本含量的方法将在本章第十节介绍。

（五）调查方法 调查分为全面调查和抽样调查两种。全面调查就是对总体的每一个个体逐一调查，其涉及的范围广、时间长、工作量大，因而需耗费大量的人力、物力和时间。

抽样调查是指在全体调查对象中，通过某种方法抽取部分的有代表性的对象作调查，并以样本去推断总体。抽样方法常用的有以下 5 种：

1、完全随机抽样 首先将有限总体内的所有个体全部编号，然后用抽签或用随机数字表的方法，随机抽取若干个个体作为样本。如欲抽样调查某猪场母猪繁殖性能，应先将母猪逐一编号，再用抽签或随机数字表按所需数量抽样，抽取的每一个体均为调查对象。完全随机抽样适用于个体均匀程度较好的总体。

2、顺序抽样　也称系统抽样或机械抽样。先将有限总体内的每个个体按其自然状态编号，然后根据调查所需的数量，按一定间隔顺序抽样。如对某牧场 500 只奶山羊进行传染性无乳症的调查，抽查 50 只。可按编号顺序每隔 10 只抽一只，但第一个调查号应从 1——10 中随机选取。此法简便易行，适用于个体分布均匀的总体。

3、分等按比例随机抽样　分等按比例随机抽样又称分层按比例随机抽样。先按某些特征或变异原因将抽样总体分成若干等次(层次)，在各等次(层次)内按其占总体的比例随机抽得各等次(层次)的样本，然后将各等次(层次)抽取的样本合并在一起即为整个调查样本。如对某地奶山羊传染性无乳症的调查，经初步了解得知，在欲调查的整个地区中，该病感染率为 80%-90%的地区占 10%，感染率为 60%-80%的地区占 60%，感染率为 20%-50%的地区占 30%。若调查 200 只山羊，则应采用按比例分等抽样，在感染率为 80%-90%的地区随机抽取 20 只，感染率为 60%-80%的地区随机抽取 120 只，感染率为 20%-50%的地区随机抽取 60 只。分等按比例随机抽样法能有效地降低抽样误差，适用于总体分布不太均匀或个体差异较大的总体。但分等不正确，会影响抽样的精确性。

4、随机群组抽样　此种抽样是把总体划分成若干个群组，然后以群组为单位随机抽样。即每次抽取的不是一个个体，而是一群动物。每次抽取的群体可大小不等，但应对被抽取群体的每一个个体逐一进行调查。随机群组抽样容易组织，节省人力、物力，适用于群体差异较大，分布不太均匀的总体。

5、多级随机抽样　当调查的总体很大、并可以系统分组时，常采用多级随机抽样的方法。例如，调查某城市奶牛 305 天的 1 胎产奶量，可采用三级抽样：农场为初级抽样单位，分场为二级抽样单位，奶牛个体为三级抽样单位。多级抽样可以估计各级的抽样误差和探讨合理的抽样方案。

（六）调查的组织工作　调查研究是一项比较复杂的工作，要动员组织大量的人力，需要一定的经费，安排一定的时间，因此，应做好人员分工、经费预算、调查进程安排、调查表的准备及调查资料的整理等项工作，如此才能保证调查研究工作有计划、有步骤地完成。一般在正式调查前，需进行预调查，以检验调查设计的可行性，并培训参予调查的工作人员，以统一标准和方法。

调查时若发现问题，应立即解决。特别要对资料进行检查，保证资料完整、正确，如发现遗漏、错误应及时补充、纠正。资料检查无误后，应妥善保存，避免丢失。

第十节　样本含量的确定

如果我们要求调查研究或试验结果精确性高，则样本含量就要大，并且越大越好。但若样本太大，就会花费过多的人力、物力和时间。特别是破坏性试验，如畜牧试验中猪、牛羊等动物的屠宰试验。即使不是破坏性试验，如在农村进行活猪体重调查时，抓猪、拴猪也容易发生掉膘现象。所以，在实际调查与试验研究中，却要求样本越小越好。但样本太小必然影响精确性。因此，需要研究在一次调查或试验中如何确定适宜样本含量的问题。

一、调查研究中样本含量的估计

（一）平均数抽样调查的样本含量估计 目前对调查研究所需样本含量，还没有一个精确的估计方法。根据以往研究，一般要求样本含量占抽样总体的 5%为最小量，对变异较小的群体，则可低于 5%。斯丹(C. Stein)认为，调查样本含量与调查要求的准确性高低及所研究对象的变异度大小有关。因此，需要提出我们能够接受的允许误差，并初步了解调查指标变异度的大小。

由标本平均数与总体平均数差异显著性检验的 t 检验公式推出的样本含量计算公式为：

$$n = t_{\alpha}^{2} S^{2} / d^{2} \qquad (12\text{-}9)$$

式中：n 为样本含量；

t_{α} 为自由度 n-1、两尾概率为 α 的临界 t 值；

S 为标准差，由经验或小型调查估得；

d 为允许误差 $(\bar{x}-\mu)$，可根据调查要求的准确性确定；

1-α 为置信度。

在首次计算时，可先用 $df=\infty$ 时 t_{α} (当置信度为 95%时，$t_{\alpha}=t_{0.05}=1.96$；置信度为 99%时，$t_{\alpha}=t_{0.01}=2.58$)值代入，若算得 $n<30$，再用 $df=n$-1 的 t_{α} 代入计算，直到 n 稳定为止。

【例 12.9】 进行南阳黄母牛体高调查，已测得南阳黄母牛的体高的标准差 S=4.07cm，今欲以 95%的置信度使调查所得的样本平均数与总体平均数的允许误差不超过 0.5cm，问需要抽取多少头黄牛组成样本才合适？

已知：S=4.07, d=0.5, 1-α =0.95，先取 $t_{0.05}$=1.96，代入(12-9)式，得：

$$n=1.96^{2}\times 4.07^{2}/0.5^{2}=254.54\approx 255 \text{ (头)}$$

即对南阳黄母牛体高进行调查，至少需要调查 255 头，才能以 95%的置信度使调查所得样本平均数与总平均数相差不超过 5cm。

（二）百分数抽样调查样本含量估计 如果我们调查的目的是对服从二项分布的总体百分数作出估计，由样本百分数与总体百分数差异显著性检验 u 检验公式推出样本含量计算公式为：

$$n=u_{\alpha}^{2} pq/d^{2} \qquad (12\text{-}10)$$

式中：n 为样本含量；

p 为总体的百分数；

q=1-p；

u_{α} 为两尾概率为 α 的临界 u 值，$u_{0.05}$=1.96，$u_{0.01}$=2.58；

d 为允许误差($\hat{p}$-p)，$\hat{p}$ 为样本百分率，可由经验得出；

1-α 为置信度。

总体百分数如果事先未知，可先从总体中调查一个样本估计。或令 p=0.5 进行估算。

【例 12.10】 欲了解某地区鸡新城疫感染率，已知道通常感染率约 60%，若规定允许误差为 3%，取置信度 1-α =0.95，问至少需要调查多少只鸡？

将 p=0.6, q=1-p=1-0.6=0.4, d=0.03, u_{α}=1.96, 代入 (12-10) 式，得：

$$n=1.96^{2}\times 0.6\times 0.4/0.03^{3}\approx 1025 \text{ (只)}$$

即至少需要调查 1025 只鸡，才能以 95%的置信度使调查所得的样本百分数与总体百分数相差不超过 0.03。

此外，当样本百分数接近 0%或 100%时，分布呈偏态，应对 x 作 $\sin\sqrt{x}$ 转换。此时估算公式为：

$$n=[57.3u_{\alpha}/\sin^{-1}(d/p\sqrt{1-p})]^2 \qquad (12\text{-}11)$$

【例 12.11】 某地需抽样调查牛结膜炎发病率，已知通常发病率为 2%，若规定允许误差为 0.1%，取置信度 1-α=0.95，问至少需要调查多少头牛？

将 p=0.02, d=0.001, u_{α}=1.96, 代入(12-11)式，得：

$$n=\{57.3\times1.96/\sin^{-1}[0.001/0.02\sqrt{(1-0.02)}]\}^2=1505\text{(头)}$$

即至少需要调查 1505 头牛，才能以 95%的置信度使估计出的牛结膜炎发病率误差不超过 0.1%。

二、试验研究中重复数的估计

（一）配对设计中重复数的估计 由配对设计 t 检验公式导出：

$$n=t_{\alpha}^{2}S_d^2/\bar{d}^2 \qquad (12\text{-}12)$$

式中：n 为试验所需动物对子数，即重复数；

S_d 为差数标准误，根据以往的试验或经验估计；

t_{α} 为自由度 n-1、两尾概率为 α 的临界 t 值；

$\bar{d}$ 为要求预期达到差异显著的平均数差值($\bar{x}_1-\bar{x}_2$)；

1-α 为置信度。

首次计算时以 $df=\infty$ 的 t_{α} 值代入计算，若 $n\leqslant15$，则以 $df=n$-1 的 t_{α} 值代入再计算，直到 n 稳定为止。

【例 12.12】 比较两个饲料配方对猪增重的影响，配对设计，希望以 95%的置信度在平均数差值达到 1.5 kg 时，测出差异显著性。根据以往经验 S_d=2 kg，问需要多少对试验家畜才能满足要求？

将 $t_{0.05(\infty)}$=1.96, S_d=2, $\bar{d}$=1.5 代入 (12-12) 式，得：

$$n=1.96^2\times2^2/1.5^2\approx7\text{（对）}$$

因为 n<15，再以 df=7-1=6 时，$t_{0.05}$=2.477 代入 (12-12) 式：

$$n=2.477^2\times2^2/1.5^2\approx11\text{（对）}$$

再以 n=11, df=11-1=10 时，$t_{0.05}$=2.2 代入 (12-12) 式：

$$n=2.2^2\times2^2/1.5^2\approx9\text{（对）}$$

再以 n=9, df=8 时，$t_{0.05}$=2.3 代入(12-12) 式：

$$n=2.3^2\times2^2/1.5^2\approx9\text{（对）}$$

n 已稳定为 9，故该配对试验至少需 9 对试验家畜才能满足试验要求。

（二）非配对试验重复数的估计 对于随机分为两组的试验，若 $n_1=n_2$，可由非配对 t 检验公式导出：

$$n = 2t_{\alpha}^{2}S/(\bar{x}_1 - \bar{x}_2)^2 \qquad (12\text{-}13)$$

式中：n 为每组试验动物头数，即重复数；

t_{α} 为 df=2(n-1)、两尾概率为 α 的临界 t 值；

S 为标准差，根据以往的试验或经验估计；

$(\bar{x}_1 - \bar{x}_2)$为预期达到差异显著的平均数差值；

1-α 为置信度。

首次计算时，以 df=∞时的 t_{α} 值代入计算，若算出的 n≤15，则以 df=2(n-1)的 t_{α} 值代入再计算，直到 n 稳定为止。

【例 12.13】 对**【例 12.12】**，若采用非配对设计，根据以往经验 S=2 *kg*，希望以 95% 的置信度在平均数差值达到 1.5 *kg* 时，测出差异显著性，问每组至少需要多少头试验家畜才能满足要求？

将 $t_{0.05(\infty)}$=1.96，S=2，$\bar{x}_1 - \bar{x}_2$=1.5 代入 (12-13) 式得：

$$n=2\times1.96^2\times2^2/1.5^2=13.66\approx14 \text{ (头)}$$

以 n=14, df=2(14-1)=26 的 $t_{0.05}$=2.056 代入(12-9) 式：

$$n=2\times2.056^2\times2^2/1.5^2=15.03\approx15 \text{ (头)}$$

再以 n=15, df=2(15-1)=28 的 $t_{0.05}$=2.048 代入 (12-9) 式：

$$n=2\times2.048^2\times2^2/1.5^2=14.91\approx15 \text{ (头)}$$

n 已稳定在 15，即本次试验两组均至少需 15 头试验家畜才能满足要求。

（三）多个处理比较试验中重复数的估计 当试验处理数 k≥3 时，各处理重复数可按误差自由度过 df_e≥12 的原则来估计。因为当 df_e 超过 12 时，F 表中的 F 值减少的幅度已很小了。

1、完全随机设计 由 df_e =k(n-1)≥12，得重复数的估算公式为：

$$n\geqslant 12/k+1 \qquad (12\text{-}14)$$

由(12-14) 式可知，若 k=3，则 n≥5；k=4，则 n≥4；……。但当处理数 k>6 时，重复数仍应不少于 3。

2、随机单位组设计 以 df_e =(k-1) (n-1)≥12，得重复数的估算公式为：

$$n\geqslant 12/(k-1)+1 \qquad (12\text{-}15)$$

由公式(12-15)可知，若 k=3，则 n≥7；k=4，则 n≥5；……。但当处理数 k>7 时，重复数仍应不少于 3。

3、拉丁方设计 若要求 df_e=(k-1) (k-2)≥12，则重复数(此时等于处理数)≥5。

所以，为了使误差自由度不小于 12，则应进行处理数(即重复数)≥5 的拉丁方试验，即进行 5×5 以上的拉丁方试验。当进行处理数为 3、4 的拉丁方试验时可将 3×3 拉丁方试验重复 6 次，4×4 拉丁方试验重复 2 次，以保证 df_e=12。

（四）两个百分数比较试验中样本含量估计 设两样本含量相等：n_1=n_2=n，n 的计算公式可由两个样本百分数差异显著性检验 u 检验公式推得：

$$n = 2u_{\alpha}^{2}\overline{pq}/\delta^2 \qquad (12\text{-}16)$$

式中：n 为每组试验的动物头数；

$\bar{p}$ 为合并百分数，由样本百分数计算，$\bar{q}=1-\bar{p}$；

δ 为预期达到差异显著的百分数差值；

u_α 为自由度等于∞、两尾概率为 α 的临界 u 值：$u_{0.05}$=1.96，$u_{0.01}$=2.58；

1-α 为置信度。

【例 12.14】 两种痢疾菌苗对鸡白痢病的免疫效果，初步试验表明，甲菌苗有效率为 22／50 = 44%，乙菌苗有效率为 28／50 = 56%，今欲以 95%的置信度在样本的百分数差值达到 10％时检验出两种菌苗免疫效果有显著差异，问试验时每组至少需接种多少只鸡？

已知 $\hat{p}_1$ =22／50 = 44%，$\hat{p}_2$ =28／50 = 56%，则两个样本百分数的合并百分数为：

$$\bar{p} = (22+28)/(50+50) = 0.50,\quad \bar{q} = 1-\bar{p} = 1-0.50 = 0.50$$

将 $u_{0.05} = 1.96, \bar{p} = 0.50, \bar{q} = 0.509, \delta = 0.10$ 代入 (12-16) 式算得：

$$n = 2\times 1.96^2 \times 0.50 \times 0.50/0.10^2 = 192.08 \approx 193(\text{只})$$

即在正式接种试验时，每组至少需接种 193 只鸡方可满足试验要求。

注意，在配对试验、非配对试验和多个处理比较试验中，同一处理的不同重复意味着同一处理实施在不同的试验单位上。若试验以个体为试验单位，则同一处理的不同重复是指同一处理实施在不同个体上；若以群体为一个试验单位，则同一处理的不同重复是指同一处理实施在不同群体上，这时如果每处理只实施在一个群体上，不管这群动物的数量有多少，实际上相当于只实施在一个试验单位上，只能获得一个观测值，也就无法估计试验误差。

习　　题

1. 动物试验的任务是什么？动物试验计划包括哪些内容？

2. 什么是试验方案？如何拟定一个正确的试验方案？

3. 产生试验误差的主要原因是什么？如何避免系统误差、降低随机误差？

4. 试验设计应遵循哪三条基本原则？这三条基本原则的相互关系与作用为何？

5. 常用的试验设计方法有哪几种？各有何优缺点？各在什么情况下应用？

6. 调查研究中常用的抽样方法有哪几种？各适用于什么情况？

7. 为了研究不同种类饲料对奶牛产奶量的影响，设置了 A、B、C、D、E 5 种饲料，用 5 头奶牛进行试验，试验根据泌乳阶段分为 5 期，每期 4 周，采用 5×5 拉丁方设计。试验结果列于下表，试对其进行方差分析。(饲料间 F=20.61)

饲料种类对奶牛产奶量影响的试验结果　　(单位：kg)

牛　号	时　　期				
	一	二	三	四	五
Ⅰ	E(300)	A(320)	B(390)	C(390)	D(380)
II	D(420)	C(390)	E(280)	B(370)	A(270)
III	B(350)	E(360)	D(400)	A(260)	C(400)
IV	A(280)	D(400)	C(390)	E(280)	B(370)
V	C(400)	B(380)	A(350)	D(430)	E(320)

8. 采用 2×2 交叉设计以研究降温对奶牛产奶量的影响。设置通风和洒水降温处理 A_1 和对照 A_2，选用胎次、产犊日期相近的泌乳中期奶牛 8 头，随机分为 B_1、B_2 两组，每组 4 头，试验分为 C_1、C_2 两期，每期 4 周，试验结果列于下表。试分析通风和洒水对产奶量有无显著影响。(F=19.86 或 t=4.57)

降温对奶牛产奶量影响的 2×2 交叉试验结果　　(千克/头·日)

时　　期		C_1	C_2
处 理		A_1	A_2
B_1 组	1	16.40	16.46
	2	19.50	14.20
	3	18.45	13.05
	4	14.15	13.55
处 理		A_2	A_1
B_2 组	1	13.75	20.10
	2	15.25	17.05
	3	15.05	18.55
	4	12.30	13.95

9. 有一多因素试验，考察因素 A、B、C、D 分别有 2 个水平，同时要考察 B 与 C 的交互作用，若用正交表 $L_8(2^7)$安排试验，请作出表头设计。

10. 为了研究粗蛋白、消化能和粗纤维三个因素对 30—50kg 育肥猪增重的影响，用正交表 $L_9(3^4)$安排了正交试验，获得下列资料。对试验结果进行方差分析。(F_A 和 F_B 均<1，F_C=1.61)

试验方案及结果表

试验号	因　　素			
	A 粗蛋白(%)	B 消化能(kJ)	C 粗纤维(%)	日增重(g)
1	1(18)	1(12970)	1(5)	475
2	1(18)	2(11715)	2(7)	394
3	1(18)	3(11460)	3(9)	362
4	2(15)	1(12970)	2(7)	445
5	2(15)	2(11715)	3(9)	392
6	2(15)	3(11460)	1(5)	409
7	3(12)	1(12970)	3(9)	354
8	3(12)	2(11715)	1(5)	378
9	3(12)	3(11460)	2(7)	423

11. 欲抽样调查某一地区仔猪断奶体重，已知 S=3.4kg，若估计断奶体重的置信度为 99%，允许误差为 0.5kg，问样本含量多少为宜？(n=308 头)

12. 某地需抽样调查猪蛔虫感染率。根据以往经验，感染率一般为 45%左右。若规定允许误差为 3.2%，选定 α =0.05，试求出样本含量。(n=929 头)

13. 某试验比较 4 个饲料配方对蛋鸡产蛋量的影响，采用随机单位组设计，若以 20 只鸡为一个试验单位，问该试验至少需要多少只鸡方可满足误差自由度不小于 12 的要求？（400 只）

附　录　　常用生物统计方法的 SAS 程序

一、SAS 系统简介

SAS 是“**Statistical Analysis System**”的缩写， 是一个用来管理分析数据和编写报告的组合软件系统。其基本部分是 SAS/BASE 软件。1966 年，美国 North Carolina 州立大学开始开发 SAS 统计软件包，1976 年该系统完成，同时成立 SAS 研究所。当初该系统只能运行于大型计算机系统，1985 年出现了当今我们广泛使用的 SAS 微机版本。SAS 系统具有统计分析方法丰富、信息储存简单、语言编程能力强、能对数据连续处理、使用简单等特点。SAS 是一个出色的统计分析系统，它汇集了大量的统计分析方法，从简单的描述统计到复杂的多变量分析，编制了大量的使用简便的统计分析过程。

二、SAS 系统运行的几个重要前提条件

（一）SAS 系统运行时要同时打开的文件较多，因此在微型计算机的系统配置文件 CONFIG.SYS 中应指定 FILES=50 或以上；

（二）SAS 系统软件有时间租期限制，因此只有机器时间（DATE）在软件有效期内才能运行。时间租期取决于 SAS 出售版本日期，即所谓的 SAS 诞生日（BIRTHDAY）。

（三）SAS 系统应全部安装到硬盘的 SAS 子目录下，硬盘应至少有 10M 空间。

三、SAS 系统的启动与关闭

（一）启动 如果 SAS 系统安装在 C 盘的子目录 SAS 下，在 WINDOWS 操作系统中，可以直接用鼠标双击桌面上 SAS 系统的快捷键图标，即进入 SAS 系统。

在 DOS 操作系统中，则开机后先进入 SAS 子目录，再输入 SAS 并按回车键即进入 SAS 显示管理系统。

C>:**cd sas**↵　　或者 C>:**cd\sas**↵

C>:\sas**sas**↵

此时屏幕上出现三个窗口，它们依次是：**OUTPUT**（SAS 结果输出窗口，在这里显示由 SAS 过程所输出的结果）、**LOG**（SAS 日志窗口，随着 SAS 语句的执行，显示出 SAS 系统的信息和已执行的语句）和 **PGM**（SAS 程序编辑窗口，在此你能输入和编辑 SAS 语句，但应注意程序不要写在行号上）。

（二）退出 SAS　在上述三个窗口的任一窗口的命令行上输入 **BYE** 或 **ENDSAS** 并按回车键即可退出 SAS。

四、SAS 程序结构、程序的输入、修改调试和运行

（一）程序结构 在 SAS 系统中任何一个完整的处理过程均可分为两大步——数据步和过程步来完成。

数据步——将不同来源的数据读入 SAS 系统建立起 SAS 数据集。每一个数据步均由 **DATA** 语句开始，以 RUN 语句结束。

过程步——调用 SAS 系统中已编号的各种过程来处理和分析数据集中的数据。每一个过程步均以 **PROC** 语句开始，RUN 语句结束，并且每个语句后均以“;”结束。

（二）程序的输入、修改调试和运行 SAS 程序只能在 PGM 窗口输入、修改，并写在 PGM 窗口预先设置好的行号区的右边。SAS 程序语句可以使用大写或小写字母或混合使用来输入，每个语句中的单词或数据项间应以空格隔开。每行输入完后加上“;”，但在数据步中 CARDS 语句后面的数据行不能加“;”，必须等到数据输入完后提行单独加“;”。在键入过程中可移动光标对错误进行修改。

SAS 语句书写格式相当自由，可在各行的任何位置开始语句的书写。一个语句可以连续写在几行中，一行中也可以同时写上几个语句，但每个语句后面必须用“;”隔开。

当一个程序输入完后，是否能运行和结果是否正确，只有将其发送到 SAS 系统中心去执行后，在 LOG 和 OUTPUT 窗口检查才能确定。发送程序的命令为 **F10** 功能键或 **SUBMIT**。当程序发送到 SAS 系统后，PGM 的程序语句全部自动清除，LOG 窗口将逐步记下程序运行的过程和出现的错误信息（用红色提示错误）。如果过程步没有错误，运行完成后，通常会在 OUTPUT 窗口打印出结果；如果程序运行出错，则需要在 PGM 窗口用 **RECALL**（或 **F9**）命令调回已发送的程序进行修改。

五、常用生物统计方法的 SAS 程序

下面结合本教材介绍几种常用生物统计分析方法的 SAS 程序，读者应注意，所提供的这些程序并不是一成不变的，根据分析的需要，每一种程序中各语句都有不同的选项，下面的程序只给出了一些最基本的语句。只要大家熟悉并掌握了 SAS 程序，就可以根据需要灵活应用。

（一）*t* 检验

1、样本平均数与总体平均数的差异显著性检验(教材【例 5.1】)

```
DATA  A;
INPUT y@@;
Y=y-114;
CARDS;
116 115 113 112 114 117 115 116 114 113
;
PROC MEANS N MEAN STDERR T PRT;
RUN;
```

程序说明：样本平均数与总体平均数的差异显著性检验可调用 MEANS 过程。DATA 语句产生临时数据集 A，表明数据步的开始；INPUT 语句指明读取变量 *y*，@@表示读入一条观测值后不换行，连续读入数据，使用@@符号可在一个物理行中输入多条观测值，减少数据输入行；CARDS 语句表明以下为数据行，数据行下的“;”表示数据行结束；PROC MEANS 语句指明调用 MEANS 过程对数据集 A 进行分析，输出样本含量 N、平均数 MEAN、平均数的标准误 STDERR、学生氏 T 值和 *t* 值概率 PRT；RUN 语句表示过程步结束，开始运行过程步。

2、配对试验资料的 *t* 检验（教材【例 5.5】）

```
DATA  B;
```

```
INPUT  ID  x1  x2;
d=x1-x2;
CARDS;
1   37.8  37.9   2   38.2  39.0   3   38.0  38.9  4   37.6  38.4
5   37.9  37.9   6   38.1  39.0   7   38.2  39.5  8   37.5  38.6
9   38.5  38.8  10   37.9  39.0
;
PROC MEANS MEAN STDERR T PRT;
   VAR d;
RUN;
```

程序说明：配对试验资料的 *t* 检验可调用 MEANS 过程。

3、非配对试验资料的 *t* 检验（教材【例 5.3】）

```
DATA  C;
INPUT  breed  y@@;
CARDS;
1  1.20  2  2.00  1  1.32  2  1.85  1  1.10  2  1.60  1  1.28  2  1.78
1  1.35  2  1.96  1  1.08  2  1.88  1  1.18  2  1.82  1  1.25  2  1.70
1  1.30  2  1.68  1  1.12  2  1.92  1  1.19  2  1.80  1  1.05;
PROC TTEST;
CLASS breed;
VAR y;
RUN;
```

程序说明：非配对试验资料的 *t* 检验需调用 TTEST 过程。INPUT 语句读入处理变量 breed（品种）和试验结果 y（增重）；CLASS 语句定义分类变量，TTEST 过程要求分类变量只能有两个水平，此处为 1（长白猪）和 2（蓝塘猪）。

（二）方差分析 对于一般的方差分析（平衡资料，即各处理重复数相等）可用 ANOVA 过程；对于非平衡资料（各处理重复数不等）的方差分析可用 GLM 过程。下面分别 ANOVA 过程和 GLM 过程。

1、ANOVA 过程的程序格式

```
PROC ANOVA 选项;
CLASS  变量;
MODEL 依变量=效应/选项;
MEANS 效应/选项;
```

程序说明：PROC ANOVA 语句中的“选项”——DATA=输入数据集，OUTSTAT=输出数据集，用于存储方差分析结果；CLASS 语句指明分类变量，此语句一定要设定，并且应出现在 MODEL 语句之前；MODEL 语句定义分析所用的线性数学模型；MEANS 语句计算各处理效应的平均数，“选项”用于设定多重比较方法——常用的有 LSD 法、DUNCAN（Duncan 新复极差法）、TUKEY（Tukey 固定极差检验法）、DUNNETT 和 DUNNETU（Dunnett 氏最小显著差数两尾和一尾检验法），显著水平的确定采用如 ALPHA=0.01（表示将显著水平设定为 0.01），缺省为 0.05。

上述语句中，关键语句在于定义线性数学模型。同一试验资料，根据模型不同而异。

常用的模型定义语句有：MODEL y=a（单因素试验资料的方差分析）、MODEL y=a b（两因素试验资料无互作模型）、MODEL y=a b c（三因素主效模型）、MODEL y=a b a*b（两因素试验资料有互作模型，也可写成 y=a|b）、MODEL y=a b(a)（两因素试验资料嵌套模型，用于系统分组资料）、MODEL y1　y2=a b（两元两因素主效模型）。

结果输出包括分类变量信息表、方差分析表和多重比较表等。

2、GLM 过程的程序格式

```
PROC ANOVA 选项;
CLASS　变量;
MODEL 依变量=效应/选项;
MEANS 效应/选项;
RANDOM 效应/选项;
CONTRAST “对比说明”效应 对比向量;
OUTPUT OUT=输出数据集 PREDICTED| P=变量名 RESIDUAL|R=变量名;
```

程序说明：PROC GLM 语句设定分析数据集和输出数据集;；CLASS 语句指明分类变量，此语句一定要设定，并且应出现在 MODEL 语句之前；MODEL 语句定义分析所用的线性数学模型和结果输出项；MEANS 语句计算平均数，并可选用多种多重比较方法；RANDOM 语句指定模型中的随机效应，“选项”——Q 给出期望均方中主效应的所有二次型；CONTRAST 语句用于对比检验；OUTPUT 语句产生输出数据集，P=定义 y 预测值变量名，R=定义误差变量名。

模型定义仍是 GLM 过程使用的关键（同上）。通过设定模型（MODEL），即可对不同的试验设计资料进行分析。当处理效应为固定效应时，通过 MEANS 语句计算平均数，进行多重比较，当处理效应为随机效应时，可利用 RANDOM 语句或 VARCOMP 过程估计方差分量。

（三）线性回归分析

1、一元线性回归分析（教材【例 8.1】）

```
DATA　G;
INPUT x y@@;
CARDS;
80　2350　86　2400　98　2720　90　2500　120　3150　102　2680
95　2630　83　2400　113　3080　105　2920　110　2960　100　2860
;
PROC REG CORR;
MODEL y=x / CLM CLI;
RUN;
```

程序说明：一元线性回归分析可调用 REG 过程。PROC 语句选项 CORR，要求输出简单相关系数；MODEL 语句指明输出 CLM——y 总体平均数的置信区间和 CLI——单个 y 值的置信区间。

2、多元线性回归分析（教材第九章习题的第 9 题）

```
DATA　H;
INPUT　number　x1　x2　x3　y@@;
CARDS;
```

```
1  23.73  5.49  1.21  15.02  2  22.34  4.32  1.35  12.62  3  28.84  5.04  1.92  14.86
4  27.67  4.72  1.49  13.98  5  20.83  5.35  1.56  15.91  6  22.27  4.27  1.50  12.47
7  27.57  5.25  1.85  15.80  8  28.01  4.62  1.51  14.32  9  24.79  4.42  1.46  13.76
10 28.96  5.30  1.66  15.18  11 25.77  4.87  1.64  14.20  12 23.17  5.80  1.90  17.07
13 28.57  5.22  1.66  15.40  14 23.52  5.18  1.98  15.94  15 21.86  4.86  1.59  14.33
16 28.95  5.18  1.37  15.11  17 24.53  4.88  1.39  13.81  18 27.65  5.02  1.66  15.58
19 27.29  5.55  1.70  15.85  20 29.07  5.26  1.82  15.28  21 32.47  5.18  1.75  16.40
22 29.65  5.08  1.70  15.02  23 22.11  4.90  1.81  15.73  24 22.43  4.65  1.82  14.75
25 20.44  5.10  1.55  14.37
;
PROC  REG  DATA=H  OUTEST=EST;
MODEL y=x1  x2  x3;
RUN;
```

程序说明：多元线性回归分析同样可调用 REG 过程。假设该数据资料被已经被建立在 A：H. DAT 标准文件中，则前面的数据步可以简化，从而直接调用 A：盘上的数据，具体程序为：

```
DATA  H; INFILE  ' A：H.DAT';
INPUT  number  x1  x2  x3  y;
PROC  REG  DATA=H  OUTEST=EST;
MODEL y=x1  x2  x3;
RUN;
```

（四）协方差分析 （教材【例 10. 1】）

```
DATA  K;
INPUT  t $  x  y@@;
CARDS;
ck  1.50  12.40  ck  1.85  12.00  ck  1.35  10.80  ck  1.45  10.00  ck  1.40  11.00
ck  1.45  11.80  ck  1.50  12.50  ck  1.55  13.40  ck  1.40  11.20  ck  1.50  11.60
ck  1.60  12.60  ck  1.70  12.50
1  1.35  10.20  1  1.20  9.40  1  1.45  12.20  1  1.20  10.30  1  1.40  11.30
1  1.30  11.40  1  1.15  12.80  1  1.30  10.90  1  1.35  11.60  1  1.15  8.50
1  1.35  12.20  1  1.20  9.30
2  1.15  10.00  2  1.10  10.60  2  1.10  10.40  2  1.05  9.20  2  1.40  13.00
2  1.45  13.50  2  1.30  13.00  2  1.70  14.80  2  1.40  12.30  2  1.45  13.20
2  1.25  12.00  2  1.30  12.80
3  1.20  12.40  3  1.00  9.80  3  1.15  11.60  3  1.10  10.60  3  1.00  9.20
3  1.45  13.90  3  1.35  12.80  3  1.15  9.30  3  1.10  9.60  3  1.20  12.40
3  1.05  11.20  3  1.10  11.00
;
PROC  GLM;
CLASS  t;
MODEL y=t  x / SOLUTION;
MEANS  t / DUNCAN;
```

LSMEANS t / STDERR PDIFF TDIFF；

RUN；

程序说明：协方差分析可调用 GLM 过程。CLASS 语句指明了分类变量为 *t*（这里代表处理，其中 ck 表示对照组，1、2、3 分别代表配方 1、配方 2、配方 3），且必须在 MODEL 语句之前。MODEL 语句定义协方差分析的数学模型。选项 SOLUTION 给出参数的估计值；MEANS 语句中，多重比较选用 DUNCAN 法（*SSR* 法）；LSMEANS 语句计算效应的最小二乘估计的平均数（LSM）；STDERR 给出 LSM 的标准误；TDIFF，FDIFF 要求显示检验 H_0：LSM（i）=LSM(j)的 *t* 值和概率值。

同前面一样，假设该数据资料被已经被建立在 A：K. DAT 标准文件中，则前面的数据步也可以简化。

主要参考文献

[1] 贵州农学院主编. 生物统计附试验设计(第二版) .农业出版社，1989。

[2] 明道绪主编. 生物统计. 中国农业科技出版社，1998。

[3] 陈善林、张 浙编著. 统计发展史. 立信会计图书用品社，1987。

[4] 方开泰、许建伦. 统计分布. 科学出版社，1987 。

[5]（美）G.W.斯奈迪格著，杨纪珂等译. 应用与农学和生物学实验的数理统计方法. 科学出版社，1964。

[6] (日)吉田 实著，关彦华等译. 畜牧试验设计. 农业出版社，1984。

[7] 明道绪主编. 兽医统计方法. 成都科技大学出版社，1991。

[8] 莫惠栋 .农业试验设计. 上海科学技术出版社，1984。

[9] 南京农业大学主编. 田间试验与统计方法 (第二版). 农业出版社，1988。

[10] （美）R·G·D·斯蒂尔， J·H·托里著， 杨纪珂等译. 数理统计的原理与方法. 科学出版社，1976。

[11] (美)S·西格尔著，北星译. 非参数统计. 科学出版社，1986。

[12] 王梓坤著. 概率论基础及其应用. 科学出版社，1976。

[13] 吴仲贤主编. 生物统计. 北京农业大学出版社，1993。

[14] 杨纪珂等. 应用生物统计. 科学出版社，1983。

[15] 俞渭江、郭卓元编著. 畜牧试验统计. 贵州科技出版社，1995。

[16] 中国科学院数学研究所统计组. 常用数理统计方法. 科学出版社，1973。

[17] 中国科学院数学研究所数理统计组. 方差分析. 科学出版社，1977。

[18] 中国科学院数学研究所数理统计组. 回归分析方法. 科学出版社，1974。

[19] 中国科学院数学研究所数理统计组. 正交试验法. 人民教育出版社，1975。

[20] 范福仁著. 生物统计学（修订本）.江苏科学技术出版社，1980。

[21] 林德光编著. 生物统计的数学原理. 辽宁人民出版社，1982。

[22] 中国科学院数学研究所概率统计室. 常用数理统计用表. 科学出版社，1974。

[23] 上海师范大学数学系概率统计教研组. 回归分析及其试验设计. 上海教育出版社，1978。

[24] (日)山田淳三著. 刘瑞三译. 统计方法的畜牧上的应用. 上海科学技术出版社，1965。

[25] 郭祖超主编. 医用数理统计方法(第三版). 人民卫生出版社，1988。

[26] 明道绪. 通径分析——显著性检验，四川农学院学报，1985。

[27] 明道绪. 通径分析的原理与方法. 农业科学导报，1986。

[28] 明道绪、刘永建. 二个处理交叉试验结果分析的 t 检验法. 四川农业大学学报，2001。

[29] 沈恒范编. 概率论讲义（第二版）.人民教育出版社，1982。

[30] 林少宫编. 基础概率与数理统计（第二版）.人民教育出版社，1978。

[31] 北京大学数学力学系数学专业概率统计组编. 正交设计. 人民教育出版社，1976。

[32] 盖钧益主编. 试验统计方法.中国农业出版社，2000。

[33] 李春喜、王文林编著. 生物统计学. 科学出版社，1997。

[34] 杨纪珂、齐翔林编著. 现代生物统计. 安徽教育出版社，1985。

[35] 刘来福、程书肖编著. 生物统计. 北京师范大学出版社，1988。

[36] 杨茂成主编. 兽医统计学. 中国展望出版社，1990。

[37] 高山林主编. 生物统计学. 中国农业出版社，1994。

[38] 杨树勤主编. 卫生统计学（第二版）.人民卫生出版社，1991。

[39] 西北农学院主编. 概率基础与数理统计. 农业出版社，1988。

[40] 上海第一医学院卫生统计教研组. 医用统计方法. 上海科学技术出版社，1979。

[41] （美）李景均著. 潘玉春、刘明孚译. 试验统计学导论. 黑龙江教育出版社，1995。

[42] 徐继初主编. 生物统计及试验设计. 农业出版社，1992。

[43] 赵仁熔、余松烈编著. 田间试验方法. 农业出版社，1979。

[44] 彭昭英著. SAS 系统应用开发指南. 北京希望电子出版社，200

[45] 董大钧主编. SAS 统计分析软件应用指南. 电子工业出版社，1993。

[46] 高惠璇、李东风、耿直等编著. SAS 系统与基础统计分析. 北京大学出版社，1995。

[47] 沈永欢等编. 实用数学手册. 科学出版社，1999。

[48] *O.N.Bishop . Statistics for Biology. 3rd .Longman Group Lincited,*1980.

[49] *S.Chattecjee and B.Price . Regression Analysis by Example. John Wiley and Soins,*1977.

[50] *T.J.Bailey. Statistical Methods in Biology. 2nd. Hodder and Stoughton,*1981.

[51] *D.C.Montogomery. Design and Analysis of Experiments. John Wiley and Soins,*1976.

[52] *Damaraju Raghavarao . Statistical Techniques in Agricultural and Biological Research. Oxford and I.B.H. Publication Co.,*1983.

[53] *W.G.Cochran. Sampling Techniques .3rd. John Wiley and Soins,*1977.

[54] *T.A.Bancroft. Topics in Intermediate Statistical Methods.Vol.1,Iowa State University Press,Ames,Iowa,U.S.A.,*1968.

[55] *Kasch.D. Biommetrie Einfuhrung in die Biostatistik. VEB Deutcher Landwirtschaftsverlag Berlin,*1983。

[56] *Weber.E. Grundriβ der biologischen Statistik. Gustav Fischer Verlag Stuttgart New Yor*k,1980。

[57] *Robert . R.S.,F.J.Rohlf. Biometry. W.H.Freeman and Company.New York,1977.*

《生物统计附试验设计》

习　题　集

第一章　　绪　论

一、名词解释

总体　个体　样本　样本含量　随机样本　　参数　　统计量　随机误差　　系统误差　准确性　精确性

二、简答题

1、什么是生物统计？它在畜牧、水产科学研究中有何作用？

2、统计分析的两个特点是什么？

3、如何提高试验的准确性与精确性？

4、如何控制、降低随机误差，避免系统误差？

第二章　资料的整理

一、名词解释

数量性状资料　　质量性状资料　半定量（等级）资料　计数资料　计量资料　全距（极差）　组中值　次数分布表　　次数分布图

二、简答题

1、资料可以分为哪几类？它们有何区别与联系？

2、为什么要对资料进行整理？对于计量资料，整理的基本步骤怎样？

3、在对计量资料进行整理时，为什么第一组的组中值以接近或等于资料中的最小值为好？

4、统计表与统计图有何用途？常用统计图有哪些？常用统计表有哪些？ 列统计表、绘统计图时，应注意什么？

第三章　　平均数、标准差与变异系数

一、名词解释

算术平均数　无偏估计　几何平均数　中位数　众数　调和平均数　标准差　方差 离均差的平方和（平方和）　　变异系数

二、简答题

1、生物统计中常用的平均数有几种？各在什么情况下应用？

2、算术平均数有哪些基本性质？

3、标准差有哪些特性？

4、为什么变异系数要与平均数、标准差配合使用？

三、计算题

1、10 头母猪第一胎的产仔数分别为：9、8、7、10、12、10、11、14、8、9 头。试计算这 10 头母猪第一胎产仔数的平均数、标准差和变异系数。

2、随机测量了某品种 120 头 6 月龄母猪的体长，经整理得到如下次数分布表。试利用加权法计算其平均数、标准差与变异系数。

组别	组中值（x）	次数（f）
80—	84	2
88—	92	10
96—	100	29
104—	108	28
112—	116	20
120—	124	15
128—	132	13
136—	140	3

3、某年某猪场发生猪瘟病，测得 10 头猪的潜伏期分别为 2、2、3、3、4、4、4、5、9、12(天)。试求潜伏期的中位数。

4、某良种羊群 1995—2000 年六个年度分别为 240、320、360、400、420、450 只，试求该良种羊群的年平均增长率。

5、某保种牛场，由于各方面原因使得保种牛群世代规模发生波动，连续 5 个世代的规模分别为：120、130、140、120、110 头。试计算平均世代规模。

6、调查甲、乙两地某品种成年母水牛的体高（*cm*）如下表，试比较两地成年母水牛体高的变异程度。

甲地	137	133	130	128	127	119	136	132
乙地	128	130	129	130	131	132	129	130

第四章　　常用概率分布

一、名词解释

必然现象　随机现象　随机试验　随机事件　概率的统计定义　小概率原理　概率分布　随机变量　离散型随机变量　连续型随机变量　概率分布密度函数　正态分布　标准正态分布　标准正态变量（标准正态离差）　双侧概率（两尾概率）　单侧概率（一尾

概率） 贝努利试验 二项分布 波松分布 返置抽样 不返置抽样 标准误 样本平均数的抽样总体 中心极限定理 *t* 分布

二、简答题

1、事件的概率具有那些基本性质？

2、离散型随机变量概率分布与连续型随机变量概率分布有何区别？

3、正态分布的密度曲线有何特点?

4、标准误与标准差有何联系与区别?

5、样本平均数抽样总体与原始总体的两个参数间有何联系？

6、*t*分布与标准正态分布有何区别与联系？

三、计算题

1、已知随机变量*u*服从N(0，1)，求P(*u*＜-1.4)， P(*u*≥1.49)， P(｜*u*｜≥2.58)， P(-1.21≤*u*＜0.45)，并作图示意。

2、已知随机变量*u*服从N(0，1)，求下列各式的u_α。

(1) P(*u*＜-u_α)+P(*u*≥u_α)=0.1;0.52

(2) P(-u_α≤*u*＜u_α)=0.42;0.95

3、猪血红蛋白含量*x*服从正态分布N(12.86，1.33^2)

(1) 求猪血红蛋白含量*x*在11.53—14.19范围内的概率。

(2) 若P(*x*＜l_1)=0.025，P(*x*＞l_2)=0.025，求l_1，l_2。

4、已知随机变量*x*服从二项分布B（100，0.1），求μ及σ。

5、已知随机变量*x*服从二项分布B(10,0.6)，求P(2≤*x*≤6)，P(*x*≥7)，P(*x*<3)。

6、已知随机变量*x*服从普阿松分布*P*(4)，求*P*(*x*=1)，*P*(*x*=2)，*P*(*x*≥4)。

第五章 *t* 检验

一、名词解释

假设检验（显著性检验） 无效假设 备择假设 显著水平 Ⅰ型错误 Ⅱ型错误 检验功效（检验力、把握度） **双侧检验**（双尾检验） **单侧检验**（单尾检验） 非配对设计（成组设计） 均数差异标准误 配对设计 自身配对 同源配对 差异标准误 *u* 检验 样本百分数标准误 参数估计 点估计 区间估计 置信区间 置信度（置信概率）

二、简答题

1、为什么在分析试验结果时需要进行显著性检验？检验的目的是什么？

2、什么是统计假设？统计假设有哪几种？各有何含义？

3、显著性检验的基本步骤是什么？根据什么确定显著水平？

4、什么是统计推断？为什么统计推断的结论有可能发生错误？有哪两类错误？如何降低两类错误？

5、双侧检验、单侧检验各在什么条件下应用？二者有何关系？

6、进行显著性检验应注意什么问题？如何理解显著性检验结论中的“差异不显著”、“差异显著”、“差异极显著”？

7、配对试验设计与非配对试验设计有何区别？

三、计算题

1、随机抽测了 10 只兔的直肠温度，其数据为：38.7、39.0、38.9、39.6、39.1、39.8、38.5、39.7、39.2、38.4（℃），已知该品种兔直肠温度的总体平均数 μ_0=39.5（℃），试检验该样本平均温度与 μ_0 是否存在显著差异？

2、11 只 60 日龄的雄鼠在 *x* 射线照射前后之体重数据见下表（单位：g）：检验雄鼠在照射 *x* 射线前后体重差异是否显著？

编　号	1	2	3	4	5	6	7	8	9	10	11
照射前	25.7	24.4	21.1	25.2	26.4	23.8	21.5	22.9	23.1	25.1	29.5
照射后	22.5	23.2	20.6	23.4	25.4	20.4	20.6	21.9	22.6	23.5	24.3

3、某猪场从 10 窝大白猪的仔猪中，每窝抽出性别相同、体重接近的仔猪 2 头，将每窝两头仔猪随机地分配到两个饲料组，进行饲料对比试验，试验时间 30 天，增重结果见下表。试检验两种饲料喂饲的仔猪平均增重差异是否显著？

窝号	1	2	3	4	5	6	7	8	9	10
饲料Ⅰ	10.0	11.2	12.1	10.5	11.1	9.8	10.8	12.5	12.0	9.9
饲料Ⅱ	9.5	10.5	11.8	9.5	12.0	8.8	9.7	11.2	11.0	9.0

4、某鸡场种蛋常年孵化率为 85%，现有 100 枚种蛋进行孵化，得小鸡 89 只，问该批种蛋的孵化结果与常年孵化率有无显著差异？

5、研究甲、乙两药对某病的治疗效果，甲药治疗病畜 70 例，治愈 53 例；乙药治疗 75 例，治愈 62 例，问两药的治愈率是否有显著差异？并计算两种药物治愈率总体百分率的 95%、99%置信区间。

第六章　方差分析

一、名词解释

方差分析　试验指标　试验因素　因素水平　试验处理　试验单位　重复　多重比较　主效应　简单效应　交互作用　　数据转换　平方根转换　对数转换　反正弦转换

二、简答题

1.多个处理平均数间的相互比较为什么不宜用 *t* 检验法?

2.方差分析在科学研究中有何意义?

3.单因素和两因素试验资料方差分析的数学模型有何区别?方差分析的基本假定是什

么?

4.进行方差分析的基本步骤为何?

5.多个平均数相互比较时，*LSD* 法与一般 *t* 检验法相比有何优点?还存在什么问题?如何决定选用哪种多重比较法?

6.为什么说两因素交叉分组单独观测值的试验设计是不完善的试验设计?在多因素试验时，如何选取最优水平组合?

7.两因素系统分组资料的方差分析与交叉分组资料的方差分析有何区别?

8.估计方差组分有何意义?

9.为什么要作数据转换?常用的数据转换方法有哪几种?各在什么条件下应用?

三、计算题

1.在同样饲养管理条件下，三个品种猪的增重如下表，试对三个品种增重差异是否显著进行检验。

品种	增重 $x_{ij}(kg)$									
A_1	16	12	18	18	13	11	15	10	17	18
A_2	10	13	11	9	16	14	8	15	13	8
A_3	11	8	13	6	7	15	9	12	10	11

2.为了比较 4 种饲料(*A*)和猪的 3 个品种(*B*)，从每个品种随机抽取 4 头猪(共 12 头)分别喂以 4 种不同饲料。随机配置，分栏饲养、位置随机排列。从 60 日龄起到 90 日龄的时期内分别测出每头猪的日增重(*g*),数据如下，试检验饲料及品种间的差异显著性。

4 种饲料 3 个品种猪 60～90 日龄日增重

	A_1	A_2	A_3	A_4
B_1	505	545	590	530
B_2	490	515	535	505
B_3	445	515	510	495

3.研究酵解作用对血糖浓度的影响，从 8 名健康人体中抽取血液并制备成血滤液。每个受试者的血滤液又可分成 4 份，然后随机地将 4 份血滤液分别放置 0、45、90、135 min 测定其血糖浓度，资料如下表。试检验不同受试者和放置不同时间的血糖浓度有无显著差异。

不同受试者、放置不同时间血滤液的血糖浓度(*mg/100ml*)

受试者编号	放置时间(*min*)			
	0	45	90	135
1	95	95	89	83
2	95	94	88	84
3	106	105	97	90
4	98	97	95	90
5	102	98	97	88
6	112	112	101	94

7	105	103	97	88
8	95	92	90	80

4.为了从 3 种不同原料和 3 种不同温度中选择使酒精产量最高的水平组合，设计了两因素试验，每一水平组合重复 4 次，结果如下表，试进行方差分析。

用不同原料及不同温度发酵的酒精产量

原料	温　度 B											
	B_1(30 ℃)				B_2(35 ℃)				B_3(40 ℃)			
A_1	41	49	23	25	11	12	25	24	6	22	26	11
A_2	47	59	50	40	43	38	33	36	8	22	18	14
A_3	48	35	53	59	55	38	47	44	30	33	26	19

5. 3 头公牛交配 6 头母牛(各随机交配两头)，其女儿第一产 305 天产奶量资料如下，试作方差分析，并估计方差组分。

公牛所配母牛的女儿产奶量(*kg*)

公牛号 S	母牛序号 D	女儿产奶量 C		母牛女儿头数	公牛女儿头数
1	1	5700	5700	2	4
	2	6900	7200	2	
2	3	5500	4900	2	4
	4	5500	7400	2	
3	5	4600	4000	2	4
	6	5300	5200	2	

6.测得某品种猪的乳头数资料列于下表。试分析公猪和母猪对仔猪乳头数的影响，并进行方差组分的估计。

某品种猪的乳头数资料

公猪 A	母猪 B	仔猪数 n_{ij}	仔猪乳头数 C				
A_1	B_{11}	8	14(3)	15(2)	16(3)		
	B_{12}	9	15(2)	16(2)	17(5)		
	B_{13}	11	12(1)	13(2)	14(5)	15(1)	16(2)
	B_{14}	10	14(2)	15(3)	16(4)	18(1)	
A_2	B_{21}	9	14(1)	15(3)	16(3)	17(1)	18(1)
	B_{22}	11	13(1)	14(2)	15(5)	16(1)	17(2)
	B_{23}	12	14(4)	15(5)	16(1)	17(1)	18(1)
	B_{24}	7	13(1)	14(2)	15(1)	16(1)	17(2)
A_3	B_{31}	8	13(2)	14(5)	15(1)		
	B_{32}	10	14(4)	15(6)			
	B_{33}	12	13(2)	14(5)	15(2)	16(3)	
合计		107					

7.　3 组小白鼠在注射某种同位素 24 小时后脾脏蛋白质中放射性测定值如下表。问芥子

气、电离辐射能否抑制该同位素进入脾脏蛋白质?(提示：先进行平方根转换，然后进行方差分析)

组别	放射性测定值(百次/*min*/*g*)									
对照组	3.8	9.0	2.5	8.2	7.1	8.0	11.5	9.0	11.0	7.9
芥子气中毒组	5.6	4.0	3.0	8.0	3.8	4.0	6.4	4.2	4.0	7.0
电离辐射组	1.5	3.8	5.5	2.0	6.0	5.1	3.3	4.0	2.1	2.7

8.用三种不同剂量的某药物治疗兔子球虫病后，粪中卵囊数的检出结果见下表。试检验三种剂量疗效差异是否显著(提示：先作对数转换 $lg(x+1)$，然后进行方差分析)。

某药物治疗兔子球虫病效果试验

剂量(*mg*/*kg*)	卵　囊　数										*n*
(Ⅰ) 15	0	0	0	0	0	0	0	0	0	0	20
	8	14	6	5	26	1	1	7	1	2	
(Ⅱ) 10	0	0	0	0	1	25	8	2	3	8	20
	22	38	5	3	50	10	28	15	2	1	
(Ⅲ) 5	220	8	30	260	96	39	86	523	47	29	20
	40	23	143	17	11	23	99	40	20	103	

9.下表为 3 组大白鼠营养试验中测得尿中氨氮的排出量。试检验各组氨氮排出量差异是否显著。(提示：先作对数转换 l*gx*，然后进行方差分析)。

组别	尿中氨氮排出量(*mg*/6 天)											
A 组	30	27	35	35	29	33	32	36	26	41	33	31
B 组	43	45	53	44	51	53	54	37	47	57	48	42
C 组	83	66	66	86	56	52	76	83	72	73	59	53

第七章　次数资料分析——χ^2 检验

一、名词解释

χ^2 分布　χ^2 的连续性矫正　适合性检验　χ^2 检验的再分割法　独立性检验

二、简答题

1. χ^2 检验与 *t* 检验、*F* 检验在应用上有什么区别？

2.适合性检验和独立性检验有何区别？

3. 什么情况下 χ^2 检验需作矫正？如何矫正？什么情况下先将各组合并后再作 χ^2 检验？合并时应注意什么问题？

4. 在什么情况下需应用 χ^2 检验的再分割法？如何对总 χ^2 值进行分割？

三、计算题

1. 两对相对性状杂交子二代 *A*—*B*—，*A*—*bb*，*aaB*—，*aabb* 4 种表现型的观察次数依次

为：315、108、101、32，问是否符合 9∶3∶3∶1 的遗传比例？

2. 某生物药品厂研制出一批新的鸡瘟疫苗，为检验其免疫力，用 200 只鸡进行试验，某中注射 100 只（经注射后患病的 10 只，不患病的 90 只），对照组（注射原疫苗组）100 只（经注射后患病的 15 只，不患病的 85 只），试问新旧疫苗的免疫力是否有差异。

3. 甲、乙、丙三个奶牛场高产奶牛、低产奶牛头数统计如下，试问三个奶牛场高、低产奶牛的构成比是否有差异。

场　地	高产奶牛	低产奶牛
甲	32	18
乙	28	26
丙	38	10

4. 对陕西三个秦川牛保种基地县进行秦川牛肉用性能外形调查，划分为优良中下 4 个等级，试问三个地区秦川牛肉用性能各级构成比差异是否显著。

地区	优	良	中	下
甲	10	10	60	10
乙	10	5	20	10
丙	5	5	23	6

第八章　直线回归与相关

一、名词解释

相关变量　回归分析　一元回归分析　多元回归分析　相关分析　简单相关分析（直线相关分析）　复相关分析　偏相关分析　样本回归系数　样本回归截距　离回归标准误　决定系数　相关系数

二、简答题

1、回归截距、回归系数与回归估计值 $\hat{y}$ 的统计意义是什么？

2、决定系数、相关系数的意义是什么？如何计算？

3、直线相关系数与回归系数的关系如何？直线相关系数与配合回归直线有何关系？

三、计算题

1、10 头育肥猪的饲料消耗（x）和增重（y）资料如下表（单位：kg），试对增重与饲料消耗进行直线回归分析，并作出回归直线。

x	191	167	194	158	200	179	178	174	170	175
y	33	11	42	24	38	44	38	37	30	35

2、试对下列资料进行直线相关和回归分析。

x	36	30	26	23	26	30	20	19	20	16
y	0.89	0.80	0.74	0.80	0.85	0.68	0.73	0.68	0.80	0.58

第九章　　*多元线性回归与多项式回归

一、名词解释

多元回归分析　多元线性回归分析　偏回归系数　　复相关系数　　最优多元线性回归方程　标准偏回归系数(通径系数)

二、简答题

1.如何建立多元线性回归方程？偏回归系数有何意义？

2.多元线性回归的显著性检验包含哪些内容？如何进行？

3.在多元线性回归分析中，如何剔除不显著的自变量？怎样重新建立多元线性回归方程？

第十章　　协方差分析

一、名词解释

协方差　协方差分析　均积

二、简答题

1、何为试验控制？如何对试验进行统计控制？

2、均积与协方差有何关系？

3、对试验进行统计控制的协方差分析的步骤有哪些?

三、计算题

1、一饲养试验，设有两种中草药饲料添加剂和对照三处理，重复9次，共有27头猪参与试验，两个月增重资料如下。由于各个处理供试猪只初始体重差异较大，试对资料进行协方差分析。

中草药饲料添加剂对猪增重试验结果表　　(单位：*kg*)

处　理	2号添加剂		1号添加剂		对照组	
观测指标	初重*x*	增重*y*	初重*x*	增重*y*	初重*x*	增重*y*
观测值	30.5	35.5	27.5	29.5	28.5	26.5
	24.5	25.0	21.5	19.5	22.5	18.5
	23.0	21.5	20.0	18.5	32.0	28.5
	20.5	20.5	22.5	24.5	19.0	18.0
	21.0	25.5	24.5	27.5	16.5	16.0
	28.5	31.5	26.0	28.5	35.0	30.5
	22.5	22.5	18.5	19.0	22.5	20.5
	18.5	20.5	28.5	31.5	15.5	16.0
	21.5	24.5	20.5	18.5	17.0	16.0

2、四种配合饲料的比较试验，每种饲料各有供试猪 10 头，供试猪的初始重（kg）及试验后的日增重（kg）列于下表，试对试验结果进行协方差分析。

处　理	Ⅰ号料		Ⅱ号料		Ⅲ号料		Ⅳ号料	
观测指标	始重 x	增重 y	始重 x	增重 y	始重 x	增重 y	始重 x	增重 y
观测值	36	0.89	28	0.64	28	0.55	32	0.52
	30	0.80	27	0.81	22	0.62	27	0.58
	26	0.74	27	0.73	26	0.58	25	0.64
	23	0.80	24	0.67	22	0.58	23	0.62
	26	0.85	25	0.77	23	0.66	27	0.54
	30	0.68	23	0.67	20	0.55	28	0.54
	20	0.73	20	0.64	22	0.60	20	0.55
	19	0.68	18	0.65	23	0.71	24	0.44
	20	0.80	17	0.59	18	0.55	19	0.51
	16	0.58	20	0.57	17	0.48	17	0.51

第十一章　非参数检验(动物医学专业用)

一、名词解释

非参数检验　符号检验　秩和检验(符号秩和检验)　等级相关分析　等级相关系数

二、简答题

1、参数检验与非参数检验有何区别？各有什么优缺点？

2、为什么在秩和检验编秩次时不同组间出现相同数据要给予“平均秩次”，而同一组的相同数据不必计算“平均秩次”？

3、两样本比较的秩和检验的检验假设是否可用 $\mu_1 = \mu_2$ 表示？为什么？

三、计算题

1、今测定了 10 头猪进食前后血糖含量变化如下表，分别用配对资料的符号检验和秩和检验法检验进食后血糖的平均含量差异是否显著？

猪号	1	2	3	4	5	6	7	8	9	10
饲前	120	110	100	130	123	127	118	130	122	145
饲后	125	125	120	131	123	129	120	129	123	140

2、将一种生物培养物以等量分别接种到两种综合培养基 *A* 和 *B* 上，共接种 10 瓶 *A* 培养基和 15 瓶 *B* 培养基。一周后计算培养壁上单位面积的生物培养物细胞平均贴壁数，获得试验数据如下：

培养基 *A*	254	140	193	153	316	473	389	257	167	147
培养基 *B*	331	257	478	339	407	396	144	357	287	568
	483	396	245	403	390					

试检验两种培养基的培养效果有无显著差异？

3、将未达到性成熟的雌性大家鼠 14 只，随机分为三组，分别为 5 只、5 只和 4 只，各注射剂量分别为 0.64 μg/鼠 1.64 μg/鼠和 2.64 μg/鼠的促性腺激素，每天一次，连续注射三天后将其杀死，取出卵巢称重。试验结果见下表：

处　　理 （注射剂量μg/鼠）	I （0.64μg/鼠）	II （1.64μg/鼠）	III （2.64μg/鼠）
卵 巢 重 量 (*mg*)	16.5	24.9	41.8
	45.0	51.6	54.6
	26.5	35.7	31.5
	32.9	33.6	39.9
	20.0	30.4	

问三种不同剂量促性腺激素对大家鼠卵巢增重效果是否有差异？

4、观察三个品种母猪乳头数得如下次数分布表。问三个品种母猪乳头数有无差异？

乳头数（个）	母猪数（头）			合　计
	品种 *A*	品种 *B*	品种 *C*	
＜12	1	2	1	4
13	3	2	2	7
14	2	5	4	11
15	4	3	2	9
16	4	6	3	13
17	3	4	2	9
＞18	2	1	2	5

5、四种抗菌素的抑菌效力比较研究，以细菌培养皿内抑菌区直径为指标，并获得如下结果：

平皿号	抗菌素Ⅰ	抗菌素Ⅱ	抗菌素Ⅲ	抗菌素Ⅳ
1	28	23	24	19
2	27	25	20	22
3	29	24	22	21
4	26	24	21	23
5	28	23	23	22

试检验四种抗菌素的抑菌效力有无显著差异？如果有显著差异，作两两比较。

6、用最佳线性无偏预测（*BLUP*）法和相对育种值（*RBV*）法对 12 头肉牛种公牛的种用价值作评定，其评定结果排序如下。问两种评定方法是否有显著相关？

序　号	1	2	3	4	5	6	7	8	9	10	11	12
BLUP 法	9 号	8 号	5 号	4 号	10 号	11 号	3 号	6 号	12 号	2 号	1 号	7 号
RBV 法	9 号	8 号	4 号	5 号	10 号	11 号	6 号	3 号	12 号	2 号	1 号	7 号

第十二章　试验设计

一、名词解释

试验设计　试验方案　完全方案　不完全方案　唯一差异原则　局部控制　完全随机设计　随机单位组设计　拉丁方　标准型拉丁方　拉丁方设计　系统抽样(机械抽样)　分等按比例随机抽样　随机群组抽样　多级随机抽样

二、简答题

1. 动物试验的任务是什么？动物试验计划包括哪些内容？

2.如何拟定一个正确的试验方案？

3. 产生试验误差的主要原因是什么？如何避免系统误差、降低随机误差？

4. 试验设计应遵循哪三条基本原则？这三条基本原则的相互关系与作用为何？

5. 常用的试验设计方法有哪几种？各有何优缺点？各在什么情况下应用？

6. 调查研究中常用的抽样方法有哪几种？各适用于什么情况？

三、计算题

1. 为了研究不同种类饲料对奶牛产奶量的影响，设置了 A、B、C、D、E 5 种饲料，用 5 头奶牛进行试验，试验根据泌乳阶段分为 5 期，每期 4 周，采用 5×5 拉丁方设计。试验结果列于下表，试对其进行方差分析。(饲料间 F=20.61)

饲料种类对奶牛产奶量影响的试验结果　　(单位：kg)

牛　号	时　　期				
	一	二	三	四	五
Ⅰ	E(300)	A(320)	B(390)	C(390)	D(380)
II	D(420)	C(390)	E(280)	B(370)	A(270)
III	B(350)	E(360)	D(400)	A(260)	C(400)
IV	A(280)	D(400)	C(390)	E(280)	B(370)
V	C(400)	B(380)	A(350)	D(430)	E(320)

2. 欲抽样调查某一地区仔猪断奶体重，已知 S=3.4kg，若估计断奶体重的置信度为 99%，允许误差为 0.5kg，问样本含量多少为宜？

3. 某地需抽样调查猪蛔虫感染率。根据以往经验，感染率一般为 45%左右。若规定允许误差为 3.2%，选定 α =0.05，试求出样本含量。

4. 某试验比较 4 个饲料配方对蛋鸡产蛋量的影响，采用随机单位组设计，若以 20 只鸡为一个试验单位，问该试验至少需要多少只鸡方可满足误差自由度不小于 12 的要求？